

# Part-based Pseudo Label Refinement for Unsupervised Person Re-identification

Yoonki Cho Woo Jae Kim Seunghoon Hong Sung-Eui Yoon  
KAIST

{yoonki, wkim97, seunghoon.hong, sungeui}@kaist.ac.kr

## Abstract

*Unsupervised person re-identification (re-ID) aims at learning discriminative representations for person retrieval from unlabeled data. Recent techniques accomplish this task by using pseudo-labels, but these labels are inherently noisy and deteriorate the accuracy. To overcome this problem, several pseudo-label refinement methods have been proposed, but they neglect the fine-grained local context essential for person re-ID. In this paper, we propose a novel Part-based Pseudo Label Refinement (PPLR) framework that reduces the label noise by employing the complementary relationship between global and part features. Specifically, we design a cross agreement score as the similarity of  $k$ -nearest neighbors between feature spaces to exploit the reliable complementary relationship. Based on the cross agreement, we refine pseudo-labels of global features by ensembling the predictions of part features, which collectively alleviate the noise in global feature clustering. We further refine pseudo-labels of part features by applying label smoothing according to the suitability of given labels for each part. Thanks to the reliable complementary information provided by the cross agreement score, our PPLR effectively reduces the influence of noisy labels and learns discriminative representations with rich local contexts. Extensive experimental results on Market-1501 and MSMT17 demonstrate the effectiveness of the proposed method over the state-of-the-art performance. The code is available at <https://github.com/yoonkicho/PPLR>.*

## 1. Introduction

Person re-identification (re-ID) aims to retrieve a person corresponding to a given query across disjoint camera views or different time stamps [59, 69]. Thanks to the discriminative power of deep neural networks, supervised approaches [20–22, 53] have achieved impressive performance in this task. Unfortunately, they require a large amount of labeled data that demands costly annotations, limiting their practicality in large-scale real-world re-ID problems. Due

to this issue, unsupervised methods that learn the discriminative features for person retrieval from unlabeled data have recently received much attention.

Prior works on unsupervised person re-ID have utilized pseudo-labels obtained by  $k$ -nearest neighbor search [25, 47, 60] or unsupervised clustering [7, 24] for training. These approaches alternate a two-stage training scheme: the label generation phase that assigns pseudo-labels and the training phase that trains a model with generated labels. Among these approaches, the clustering-based methods [2, 9] have especially demonstrated their effectiveness with state-of-the-art performance. However, inherent noises in pseudo-labels significantly hinder the performance of these unsupervised methods.

To tackle this problem, many efforts have been made to improve the accuracy of pseudo-labels by performing robust clustering [9, 62] or pseudo-label refinement [25, 64]. Recent techniques [8, 63] significantly reduce the label noise through the model ensemble in a peer-teaching manner by using predictions from an auxiliary network as refined labels for the target network. Nevertheless, training multiple backbones as teacher networks (e.g., dual ResNet in MMT [8], and single DenseNet, ResNet, and Inception-v3 in MEB-Net [63]) requires high computational costs. Furthermore, labels refined by these methods consider only global features and neglect the fine-grained clues essential to person re-ID, leading to insufficient performance.

To address the aforementioned problems, we propose *Part-based Pseudo Label Refinement (PPLR)*, a novel unsupervised re-ID framework that effectively handles the label noise using part features in a self-teaching manner. Several studies [42, 67] demonstrate that the fine-grained information from part features improves the re-ID performance. Our key idea is that this fine-grained information can provide not only useful cues for better representation learning but also robustness against label noises. In contrast to the global-shape information that has large variations due to significant changes in poses and viewpoints, part features can capture the local-texture information that provides a more crucial clue to re-identifying a person [70].

We argue that the complementary relationship between

the global and part features can be used to refine the label noise in each of their feature spaces. However, some of the global and part features from the same image capture very different semantic information, and using the complementary relationship naively can result in noisy and even incorrect information. For instance, images may contain irrelevant parts (*e.g.*, occlusions or backgrounds) that provide unreliable complementary information, and it is desirable to exclude them from the training. Therefore, it is essential to identify whether the information of global and part features are reliable with each other to properly exploit their complementary relationship. To address this issue, we design a cross agreement score based on the similarity between the  $k$ -nearest neighbors of global and part features. Based on the cross agreement, we propose two pseudo-label refinement methods – *part-guided label refinement (PGLR)* and *agreement-aware label smoothing (AALS)*. PGLR refines the pseudo-labels of global features by aggregating the predictions of part features, guiding the global features to learn from rich local contexts. AALS refines the pseudo-labels of part features by smoothing the label distributions, thus calibrating the predictions of part features.

Our contributions can be summarized as follows:

- We propose a part-based pseudo-label refinement framework that operates in a self-ensemble manner without auxiliary networks. To the best of our knowledge, this is the first work to handle the label noise using the part feature information for person re-ID.
- We design a cross agreement score to capture reliable complementary information, which is computed by the similarity between the  $k$ -nearest neighbors of the global and part features.
- Extensive experimental results with superior performance against the state-of-the-art methods demonstrate the effectiveness of the proposed method.

## 2. Related Work

**Learning with noisy labels.** Due to the difficulty of obtaining high-quality labels in many real-world scenarios, much attention has been given to robust training with noisy labels [38]. Loss adjustment approaches utilize the loss correction technique through the noise transition matrix [13, 30, 45, 54] or the sample re-weighting scheme based on the reliability of a given label [1, 32, 36] to reduce the influence of noisy labels. However, these methods require a certain number of clean labels to estimate the degree of noise and are not applicable for unsupervised person re-ID, where the pseudo-labels are extremely noisy at the beginning of training. There also have been attempts to design a robust loss function against the label noise. Ghosh *et al.* [10] demonstrate that mean absolute error (MAE) loss is theoretically robust to noisy labels. Generalized cross-entropy (GCE) [65] loss was proposed to achieve both the

robustness of MAE and better convergence of the cross-entropy loss. Wang *et al.* [51] propose symmetric cross-entropy (SCE) loss which boosts the noise tolerance with the reverse cross-entropy loss. These loss functions, however, are designed for simple image classification tasks and are not suitable for open-set person re-ID tasks.

**Part-based approaches for person re-ID.** Fine-grained information on human body parts is an essential clue to distinguish people, and recent studies [40, 53, 66] leverage part features and show state-of-the-art performance. There have been many efforts to learn more discriminative part features through human parsing [11, 17], attention mechanism [22, 37, 57], pose estimation [29, 41], and multiple granularities [48, 67]. Despite the remarkable achievements of part-based supervised approaches, there have been only a few attempts to utilize part features for unsupervised person re-ID. For domain adaptive re-ID, PAUL [58] extracts part features using a spatial transformer network [16] for patch-based discriminative learning. Recent methods [7, 25] utilize part features to exploit robust feature similarity for accurate pseudo-labels. Contrary to the above methods, our work utilizes part features to reduce the label noise of global feature clustering by providing fine-grained information.

**Unsupervised approaches for person re-ID.** The existing unsupervised approaches can be divided into unsupervised domain adaptation (UDA) and unsupervised learning depending on whether an external labeled source domain data is used. Several UDA methods reduce the domain gap between the source and target datasets through feature distribution alignment [23, 49] and image style transfers [5, 52, 73]. In recent years, both UDA and unsupervised learning methods leverage pseudo-labels assigned by clustering [24, 39, 61] or nearest neighbor search [47, 60, 74]. Recent clustering-based methods [2, 9, 50] apply the contrastive learning scheme with cluster proxies and show impressive achievements. However, the inherent label noise in pseudo-labels degrades the performance, and many recent studies tackle this essential problem. SSL [25] softens pseudo-labels by measuring the robust similarity with additional information such as camera labels. MMT [8] and MEBNet [63] refine pseudo-labels using the predictions of auxiliary teacher networks in a mutual learning manner. RLCC [64] utilizes a label propagation scheme based on the clustering consensus matrix to reduce the label noise. Unlike these label refinement methods that only consider the global context, our work employs fine-grained information from the part features to refine the labels more effectively.

## 3. Method

We propose a Part-based Pseudo Label Refinement (PPLR) framework that exploits the complementary relationship between the global and part features to tackle the

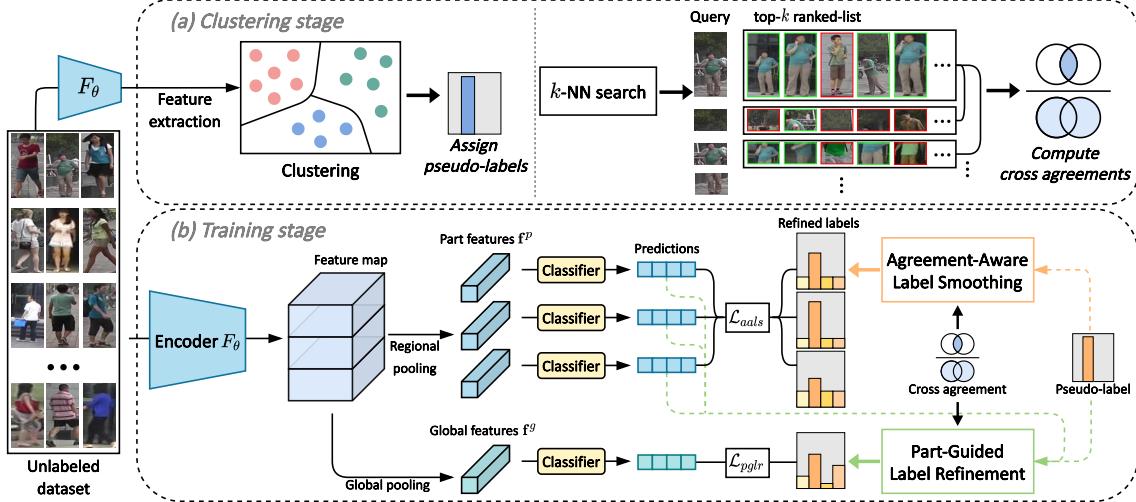


Figure 1. The illustration of PPLR. Our method alternates the clustering stage and the training stage. (a) In the clustering stage, we assign pseudo-labels by clustering the global features on the unlabeled dataset. We then perform a  $k$ -nearest neighbor search on each feature space and compute the cross agreement score based on the similarity between the top- $k$  ranked lists of the global and part features. (b) In the training stage, we train the model with refined pseudo-labels based on the cross agreement score. We smooth the labels of part features according to the cross agreement score of each part and refine the labels of global features by aggregating the part features’ predictions.

label noise problem. Following the existing clustering-based methods [24, 61, 64], our method alternates the clustering stage and the training stage. In the clustering stage, we extract global and part features and assign pseudo-labels through global feature clustering. We then compute a cross agreement score for each sample based on the similarity between  $k$ -nearest neighbors of global and part features. In the training stage, we mitigate the label noise using the proposed pseudo-label refinement methods based on the cross agreement: agreement-aware label smoothing (AALS) for part features and part-guided label refinement (PGLR) for global features. The features from the trained model are then used in the next clustering stage to update the pseudo-labels. The overall framework is illustrated in Fig. 1.

### 3.1. Part-based Unsupervised re-ID Framework

We first present a part-based unsupervised person re-ID framework that utilizes fine-grained information of the part features. Contrary to most existing unsupervised approaches that exploit only the global feature, we use both the global and part features to represent an image.

Formally, let  $\mathcal{D} = \{x_i\}_{i=1}^{N_D}$  denote the unlabeled training dataset, where  $x_i$  is an image and  $N_D$  is the number of images. Our model first extracts the shared representation  $F_\theta(x_i) \in \mathbb{R}^{C \times H \times W}$ , where  $C$ ,  $H$ , and  $W$  are sizes of the channel, height, and width of the feature map, respectively. Given this feature map, the global feature  $\mathbf{f}_i^g$  is obtained by applying global average pooling over the feature map, while the part features  $\{\mathbf{f}_i^{p_n}\}_{n=1}^{N_p}$  are obtained by dividing the feature map into  $N_p$  uniformly partitioned regions  $\mathbb{R}^{C \times \frac{H}{N_p} \times W}$  and applying average pooling on each region.

To learn these representations without a label, we simulate the pseudo-labels based on clustering results. Following the part-based approaches [42, 48, 67] in the literature, we adopt the standard protocol where both the global and part features share the same pseudo-labels. We perform DBSCAN clustering [6] on the global feature set  $\{\mathbf{f}_i^g\}_{i=1}^{N_D}$  and use the cluster assignment as pseudo-labels. We denote the pseudo-label for the image  $x_i$  as  $y_i \in \mathbb{R}^K$ , which is the one-hot encoding of the hard assignment with  $K$  clusters.

The pseudo-labels are then used to train the global and part features for person identification. For global features, we compute the cross-entropy loss by:

$$\mathcal{L}_{gce} = - \sum_{i=1}^{N_D} y_i \cdot \log(q_i^g), \quad (1)$$

where  $q_i^g = h_{\phi_g}(\mathbf{f}_i^g) \in \mathbb{R}^K$  is the prediction vector by the global feature, and  $h_{\phi_g}(\cdot)$  is the global feature classifier consisting of a fully connected layer and a softmax function. Similarly, we train the part features using the cross-entropy loss by:

$$\mathcal{L}_{pce} = - \frac{1}{N_p} \sum_{i=1}^{N_D} \sum_{n=1}^{N_p} y_i \cdot \log(q_i^{p_n}), \quad (2)$$

where  $q_i^{p_n} = h_{\phi_{p_n}}(\mathbf{f}_i^{p_n}) \in \mathbb{R}^K$  indicates the prediction vector by the  $n$ -th part feature space  $p_n$ , and  $h_{\phi_{p_n}}$  is the classifier for the part feature space  $p_n$ . We additionally utilize the softmax-triplet loss defined by:

$$\mathcal{L}_{triplet} = - \sum_{i=1}^{N_D} \log \left( \frac{e^{\|\mathbf{f}_i^g - \mathbf{f}_{i,n}^g\|}}{e^{\|\mathbf{f}_i^g - \mathbf{f}_{i,p}\|} + e^{\|\mathbf{f}_i^g - \mathbf{f}_{i,n}^g\|}} \right), \quad (3)$$

where  $\|\cdot\|$  denotes the  $L_2$ -norm, and the subscripts  $(i, p)$  and  $(i, n)$  respectively denote the hardest positive and negative samples of the image  $x_i$  in a mini-batch obtained by the hard-batch triplet selection [14]. Following the recent studies [2, 55] that utilize camera labels to improve the discriminability across camera views, we can optionally employ a camera-aware proxy [50] if the camera labels are available. We compute the camera-aware proxy  $\mathbf{c}_{(a,b)}$  as the centroid of the features that have the same camera label  $a$  and belong to the same cluster  $b$ . We then compute the inter-camera contrastive loss [50] as:

$$\mathcal{L}_{cam} = - \sum_{i=1}^{N_D} \frac{1}{|P_i|} \sum_{j \in \mathcal{P}_i} \log \frac{\exp(\mathbf{c}_j^\top \mathbf{f}_i^g / \tau)}{\sum_{k \in \mathcal{P}_i \cup \mathcal{Q}_i} \exp(\mathbf{c}_k^\top \mathbf{f}_i^g / \tau)}, \quad (4)$$

where  $\mathcal{P}_i$  and  $\mathcal{Q}_i$  are the index sets of the positive and hard negative camera-aware proxies<sup>1</sup> for  $\mathbf{f}_i^g$ , and  $\tau$  is the temperature parameter. This loss pulls together the proxies that are within the same cluster but in different cameras, reducing the intra-class variance caused by disjoint camera views. The training objective is then given by:

$$\mathcal{L} = \mathcal{L}_{gce} + \mathcal{L}_{pce} + \mathcal{L}_{triplet} + \lambda_{cam} \mathcal{L}_{cam}, \quad (5)$$

where  $\lambda_{cam}$  is the weight parameter that controls the importance of the inter-camera contrastive loss.

Ideally, the model can learn both the holistic and local features by sharing the common representation  $F_\theta(\cdot)$ . However, its performance is inherently bounded by the quality of the pseudo-label  $y_i$ , which is significantly noisy in practice. In the following sections, we propose a method to refine such noisy labels for properly representing both features.

### 3.2. Cross Agreement

Contrary to our basic framework, PPLR trains the model with refined pseudo-labels that consider the complementary relationship between global and part features. Nevertheless, there exists unreliable complementary information due to differences in feature similarity structures between global and part features. As shown in Fig. 2, some part features contain information irrelevant to a person and are not suitable for refining pseudo-labels of global features. Furthermore, global features consider only the global context and sometimes neglect information relevant to part features. Therefore, identifying whether the given complementary information is reliable is an essential task for our method.

To address this issue, we design a cross agreement score that captures how reciprocally similar the  $k$ -nearest neighbors of global and part features are. We define the cross agreement score as the Jaccard similarity between the  $k$ -nearest neighbors of the global and part features. We first perform a  $k$ -nearest neighbor search on the global and each

<sup>1</sup>More details can be found in the appendix.

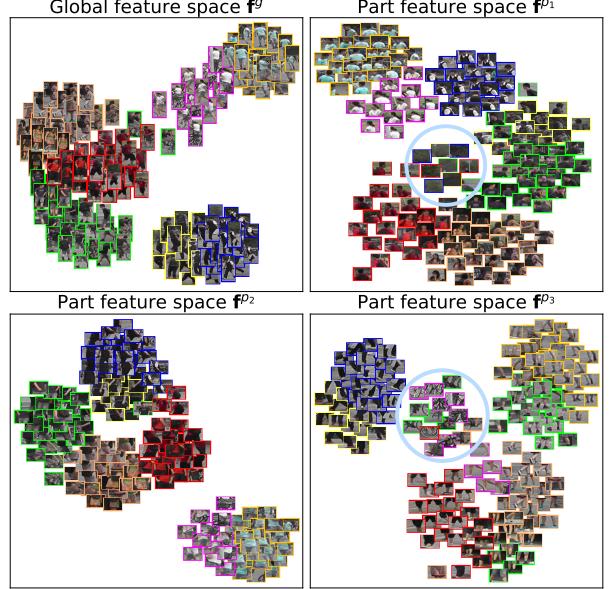


Figure 2. The t-SNE [46] visualization of each feature space on Market-1501 at an early training epoch. Different bounding box colors represent different IDs. Each feature space shows a different feature distribution with different semantic parts of a person, and some feature information can be unreliable. For instance, less discriminative parts denoted by a circle provide irrelevant information to their counterparts in other feature spaces and vice-versa. Our cross agreement score identifies such noisy information by comparing the  $k$ -nearest neighbors between feature spaces.

of the part feature spaces independently to produce  $(1+N_p)$  ranked lists on each image. We then compute the cross agreement score between the global feature space  $g$  and the  $n$ -th part feature space  $p_n$  for the image  $x_i$  by:

$$\mathcal{C}_i(g, p_n) = \frac{|\mathcal{R}_i(g, k) \cap \mathcal{R}_i(p_n, k)|}{|\mathcal{R}_i(g, k) \cup \mathcal{R}_i(p_n, k)|} \in [0, 1], \quad (6)$$

where  $\mathcal{R}_i(g, k)$  and  $\mathcal{R}_i(p_n, k)$  are the sets of indices for the top- $k$  samples in the ranked list computed by  $\mathbf{f}_i^g$  and  $\mathbf{f}_i^{p_n}$ , respectively, and  $|\cdot|$  is the cardinality of a set.

Intuitively, a high cross agreement score  $\mathcal{C}_i(g, p_n)$  implies that the feature spaces of  $g$  and  $p_n$  have highly correlated feature similarity structure around the data point  $i$  and provide reliable complementary information. On the other hand, a low  $\mathcal{C}_i(g, p_n)$  implies that the global and part features are not very correlated, meaning that they can provide unreliable information to each other. Our cross agreement score is designed in the same spirit as recent re-ranking techniques [15, 34, 56, 71] that utilize a reciprocity check of  $k$ -nearest neighbors to handle the similarity noise in the affinity matrix to improve retrieval performance.

### 3.3. Pseudo Label Refinement

Based on the cross agreement scores, we alleviate the label noise by considering (1) whether the pseudo-labels by

global feature clustering are suitable for each part feature and (2) whether the predictions of part features are appropriate for refining pseudo-labels of global features.

**Agreement-aware label smoothing.** Learning all part features with the same global pseudo-label that neglects the local context of parts can be detrimental to the model training. For instance, some parts contain cues irrelevant to a person (*e.g.*, occlusions), and it is desirable to exclude them from the training. To address this issue, we utilize a label smoothing [27, 43] to refine the pseudo-label of each part depending on the corresponding cross agreement score.

Given the pseudo-label  $y_i$  of the image  $x_i$ , the label smoothing for the part feature  $\mathbf{f}_i^{p_n}$  is formulated as:

$$\tilde{y}_i^{p_n} = \alpha_i^{p_n} y_i + (1 - \alpha_i^{p_n}) u, \quad (7)$$

where  $u$  is a uniform vector, and  $\alpha_i^{p_n}$  is a weight determining the strength of label smoothing. Contrary to conventional label smoothing that employs a constant weight for  $\alpha_i^{p_n}$ , we dynamically adjust the weight for each part using the cross agreement score (*i.e.*,  $\alpha_i^{p_n} = \mathcal{C}_i(g, p_n)$ ) that reflects the reliability of the global clustering result for each part. We then plug the refined pseudo-labels  $\tilde{y}_i^{p_n}$  to Eq. (2), and the cross-entropy loss is reformulated with Kullback-Leibler (KL) divergence [31] by:

$$\begin{aligned} \mathcal{L}_{aals} = & \frac{1}{N_p} \sum_{i=1}^{N_D} \sum_{n=1}^{N_p} (\alpha_i^{p_n} H(y_i, q_i^{p_n}) \\ & + (1 - \alpha_i^{p_n}) D_{\text{KL}}(u \| q_i^{p_n})), \end{aligned} \quad (8)$$

where  $H(\cdot, \cdot)$  and  $D_{\text{KL}}(\cdot \| \cdot)$  are cross-entropy and KL divergence, respectively, and two terms are balanced by  $\alpha_i^{p_n}$  with the value of the cross agreement score  $\mathcal{C}_i(g, p_n)$ .

In Eq. (8), the former term drives the prediction to high confidence close to  $y_i$ , and the latter term encourages the prediction to collapse into a uniform vector. By scaling the two opposite terms with the cross agreement scores, we calibrate the prediction of part features according to the reliability of pseudo-labels for each part.

**Part-guided label refinement.** We propose a part-guided label refinement that generates refined labels for global features using the predictions by part features. The part feature information with rich local contexts can be used to handle the label noise in global feature clustering, which often neglects fine-grained information. However, since less discriminative parts can provide misleading information, we aggregate the predictions of part features with different weights depending on each cross agreement score, thus refining the labels with more reliable information.

Specifically, we generate the part-guided refined label,  $\tilde{y}_i^g$ , as a pseudo-label for the global feature by:

$$\tilde{y}_i^g = \beta y_i + (1 - \beta) \sum_{n=1}^{N_p} w_i^{p_n} q_i^{p_n}, \quad (9)$$

where  $w_i^{p_n} = \frac{\exp(\mathcal{C}_i(g, p_n))}{\sum_k \exp(\mathcal{C}_i(g, p_k))}$  and  $q_i^{p_n}$  are the ensemble weight and the prediction vector of the part feature  $\mathbf{f}_i^{p_n}$ , respectively.  $\beta \in [0, 1]$  is the weighting parameter controlling the ratio of the one-hot pseudo-label and the ensembled prediction. Contrary to the global feature that only captures holistic characteristics of a person, the part-guided refined label in Eq. (9) additionally considers the fine-grained predictions from the local parts in proportion to their reliability captured by the cross agreement score. The refined labels  $\tilde{y}_i^g$  are then plugged to Eq. (1) to train the global feature by:

$$\mathcal{L}_{pglr} = - \sum_{i=1}^{N_D} \tilde{y}_i^g \cdot \log(q_i^g). \quad (10)$$

With the part-guided refined labels, global features learn from the ensembled part predictions with rich fine-grained information that is neglected in previous methods. Furthermore, unlike previous studies [8, 63] that refine pseudo-labels using an auxiliary teacher network, a part-guided label refinement is a self-teaching method without an additional network, being computationally efficient.

**Overall training objective.** The overall loss function of PPLR is then:

$$\mathcal{L}_{PPLR} = \mathcal{L}_{aals} + \mathcal{L}_{pglr} + \mathcal{L}_{triplet} + \lambda_{cam} \mathcal{L}_{cam}. \quad (11)$$

Our method effectively reduces the influence of noisy labels in two ways. The part features with low cross agreements are trained by pseudo-labels close to a uniform distribution by Eq. (8), and the global features trained by the part-guided refined labels capture reliable fine-grained information from the part features by Eq. (9). Also, when all part predictions have low cross agreement scores, the ensembled prediction in the part-guided refined label eventually collapses to a uniform vector due to the strong label smoothing effect in all parts, thus providing meaningless training signals. It allows us to weaken the impact of noisy pseudo-labels, resulting in better representation learning.

## 4. Experiments

### 4.1. Datasets and Evaluation Protocols

We evaluate the proposed method on two person re-ID datasets and a vehicle re-ID dataset – Market-1501 [68], MSMT17 [52], and VeRi-776 [26]. Market-1501 contains 32,688 images of 1,501 person identities from 6 non-overlapping camera views. It is split into 12,936 training images of 751 identities and 19,732 testing images of 750 identities. MSMT17 is a more challenging dataset consisting of 126,441 images of 4,101 person identities captured from 15 different cameras. It is split into 32,621 training images of 1,041 identities and 93,820 testing images of 3,060 identities. VeRi-776 contains 51,035 images of 776 vehicles from 20 camera views. It is split into 37,778 training

Method	Market-1501		MSMT17	
	mAP	Rank-1	mAP	Rank-1
Baseline1 (w/o $\mathcal{L}_{cam}$ )	73.5	88.5	25.1	51.2
+ $\mathcal{L}_{aals}$	78.9	90.9	28.8	56.1
+ $\mathcal{L}_{pglr}$	77.8	90.7	29.3	55.6
+ $\mathcal{L}_{aals}$ + $\mathcal{L}_{pglr}$	<b>81.5</b>	<b>92.8</b>	<b>31.4</b>	<b>61.1</b>
Baseline2 (w/ $\mathcal{L}_{cam}$ )	77.9	91.2	36.8	67.6
+ $\mathcal{L}_{aals}$	82.4	93.2	40.5	71.2
+ $\mathcal{L}_{pglr}$	81.5	93.0	40.4	70.9
+ $\mathcal{L}_{aals}$ + $\mathcal{L}_{pglr}$	<b>84.4</b>	<b>94.3</b>	<b>42.2</b>	<b>73.3</b>

Table 1. Ablation study on individual components of PPLR.

images of 576 vehicles and 13,257 testing images of 200 vehicles. We adopt mean average precision (mAP) and cumulative matching characteristic (CMC) Rank-1, Rank-5, Rank-10 accuracies to evaluate performances.

**Implementation Details.** We adopt ResNet-50 [12] pre-trained on ImageNet [4] as the backbone. We remove all layers after layer-4 and add average pooling layers followed by fully-connected classifiers with BNNeck [28]. During testing, we use only global features for retrieval. The person and vehicle images are resized to  $384 \times 128$  and  $256 \times 256$ , respectively. Random flipping, cropping, and erasing [72] are used for data augmentation. The mini-batch size is 64 consisting of 16 pseudo-classes and 4 images for each class. Adam [18] with weight decay of  $5 \times 10^{-4}$  is adopted for training. The initial learning rate is set to  $3.5 \times 10^{-4}$  and is decreased by a factor of 10 after every 20 epochs. We train for a total of 50 epochs, and each epoch contains 400 iterations. We employ DBSCAN [6] based on Jaccard distance with  $k$ -reciprocal encoding [71] for clustering. We empirically set the number of parts  $N_p$  to 3, the weighting parameter  $\beta$  to 0.5, and the parameter  $k$  of cross agreement scores to 20. Following CAP [50], we set  $\tau = 0.07$ ,  $\lambda_{cam} = 0.5$ , and the number of hard negative proxies to 50.

## 4.2. Ablation Study

To analyze the effectiveness of our method, we conduct extensive experiments on Market-1501 and MSMT17. Since camera labels are not always available in practice, we adopt the part-based unsupervised re-ID framework (Sec. 3.1) with and without the inter-camera contrastive loss  $\mathcal{L}_{cam}$  as the baselines. We evaluate the effectiveness of the proposed pseudo-label refinement methods – agreement-aware label smoothing  $\mathcal{L}_{aals}$  and part-guided label refinement  $\mathcal{L}_{pglr}$ . The experimental results on each setting are reported in Table 1. As shown in the table, each label refinement significantly boosts performances on both settings. We especially obtain remarkable performance gains when we combine AALS and PGLR. For instance, our method improves mAP of the baselines with and without  $\mathcal{L}_{cam}$  by 6.5% and 8.0%, respectively, on Market-1501.

**Effectiveness of agreement-aware label smoothing.** To verify the necessity of AALS, we evaluate label refinement techniques alternative to AALS. We first adopt vanilla la-

Method	Market-1501		MSMT17	
	mAP	Rank-1	mAP	Rank-1
PPLR	<b>81.5</b>	<b>92.8</b>	<b>31.4</b>	<b>61.1</b>
PPLR w/o $\mathcal{L}_{aals}$	77.8	90.7	29.3	55.6
+ Part-to-part label refinement	78.6	91.1	27.9	54.9
+ Vanilla label smoothing	79.1	91.7	29.8	57.2
PPLR w/o $\mathcal{L}_{pglr}$	78.9	90.9	28.8	56.1
+ Mean-teaching	79.0	91.0	28.9	55.7
+ $\mathcal{L}_{pglr}$ w/o cross agreement	80.1	91.5	30.1	58.1

Table 2. Ablation study on different label refinement techniques.

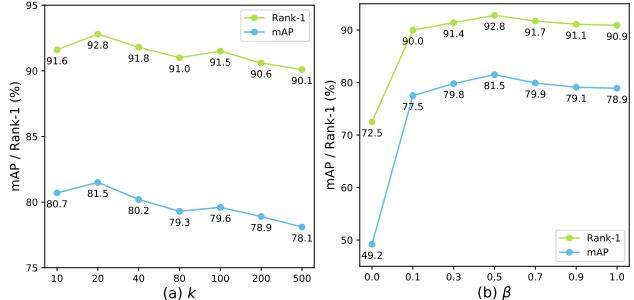


Figure 3. Parameter analysis of  $k$  and  $\beta$  on Market-1501.

bel smoothing that employs a constant smoothing weight in Eq. (7). We also explore part-to-part label refinement, which generates refined labels using the predictions of other parts in the same way as PGLR in Eq. (9). Specifically, we generate refined labels for the part  $p_i$  by aggregating the predictions of the other parts  $p_j$  with part-to-part cross agreement scores  $C_i(p_i, p_j)$ . As shown in Table 2, our AALS significantly outperforms other label refinement methods. We observe that the part-to-part cross agreement score is substantially lower than the part-to-global counterpart because the parts share less overlapped receptive fields. Thus, part-to-part label refinement is guided by unreliable information from significantly different local contexts, while our AALS captures reliable complementary information and shows better performance. Vanilla label smoothing adjusts the label distribution for all parts blindly without considering the characteristics of each part and also shows limited improvement. Meanwhile, our AALS calibrates the part features’ predictions according to the reliability of given labels captured by the cross agreements, and the well-calibrated part predictions lead to more reliable part-guided refined labels.

**Effectiveness of part-guided label refinement.** To verify the effectiveness of PGLR, we evaluate other label refinement techniques. One way is to refine labels with the prediction of global features by the mean-teacher model [44]. We further investigate PGLR without cross agreement scores by averaging the predictions of part features, *i.e.* set all  $w_i^{pn}$  to  $1/N_p$  in Eq. (9). As shown in Table 2, our PGLR significantly outperforms other label refinement methods. It demonstrates the superiority of PGLR and the effectiveness of the cross agreement score. The refined pseudo-label by PGLR captures reliable fine-grained information that cannot be achieved by considering only global features, and it helps to generate more effective refined labels.

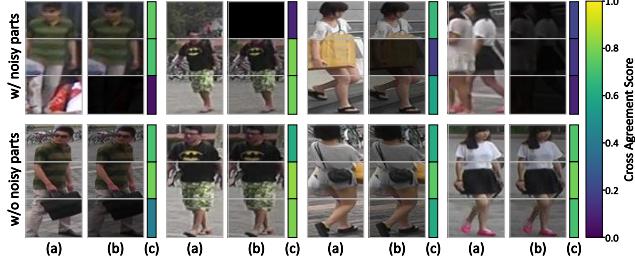


Figure 4. Visualization of cross agreement scores. (a) Original images. (b) Masked images created by multiplying the image intensity with the corresponding cross agreement score. (c) The color jet bar of the cross agreement score, where the color of each region indicates the average score over the entire training epochs.

**Parameter analysis.** We analyze the impact of two parameters in our method – the number of  $k$ -nearest neighbors for cross agreement scores and the weighting parameter  $\beta$  for the part-guided label refinement. We tune the value of each parameter while keeping the others fixed, and the results are in Fig. 3. Large  $k$  values result in more frequent false matches in top- $k$  ranked lists of global and part features, producing lower cross agreement scores overall. These false matches prevent us from identifying the reliability of complementary relationships, thus limiting performance. When we set  $\beta$  to 0, our method decomposes down to using only the ensembled part predictions in Eq. (9), showing the significant performance drop. The predictions of the initial training stage usually output uniform distributions, so the labels refined by PGLR also collapse to uniform distributions, providing noisy training signals. Based on these experimental results, we set  $k = 20$  and  $\beta = 0.5$ .

**Qualitative analysis.** To explore the effect of cross agreement scores, we visualize the masked image  $I'(x, y) = I(x, y) \circ c(x, y)$  as the element-wise product between the intensity of the image  $I(x, y)$  and the corresponding cross agreement score  $c(x, y)$ . A low cross agreement score leads to a low pixel intensity; thus, the corresponding part of an image becomes dark. As shown in Fig. 4, occluded or misaligned parts have relatively low cross agreement scores while discriminative parts have high values. In the rightmost case on the top row of Fig. 4, all parts fail to capture discriminative information due to occlusions by several people, and cross agreement values for all three parts are low, collapsing predictions of all parts to a uniform distribution by Eq. (8). The ensembled prediction of part-guided refined labels in Eq. (9) also collapses to a uniform vector, providing meaningless training signals.

To further explore the proposed label refinement methods, we qualitatively analyze the effectiveness of AALS and PGLR. We visualize the embeddings of the part feature of our PPLR with and without AALS. As shown in Fig. 5, the embeddings without AALS are overfitted to the ID labels and show less discrimination over less informa-

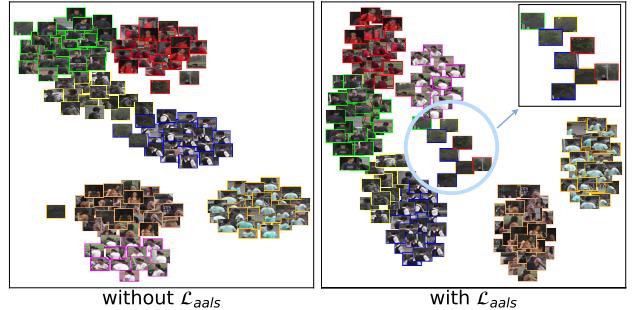


Figure 5. The t-SNE [46] visualization of the topmost part feature space of our PPLR without and with AALS. Without AALS, the embeddings overfit to ID labels and consider less the local context of each part. With AALS, it shows reliable embedding results that are well-distributed while considering the local context. The circled region shows that the features with less discriminative parts are embedded together even though their ID labels are different.

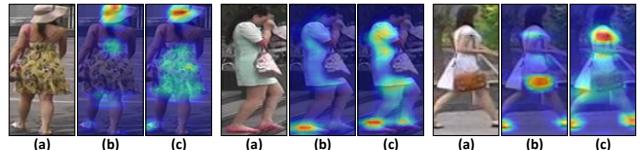


Figure 6. Grad-CAM [35] visualization of the global features’ predictions: (a) Original images; (b) without  $\mathcal{L}_{pglr}$ ; (c) with  $\mathcal{L}_{pglr}$ . (b) Without PGLR, the model focuses only on the most discriminative parts in the global context. (c) Thanks to the part-guided refined labels with rich local contexts, the model focuses on more diverse discriminative regions.

tive parts. On the other hand, embeddings with AALS consider the semantic context of each part and are better distributed. While the model without AALS relies on hard pseudo-labels, AALS adjusts the given pseudo-labels based on the reliability of each part and achieves more proper part representation learning with a calibration effect.

We also visualize Grad-CAM [35] of the predictions by global features with and without PGLR. As shown in Fig. 6, the former focuses only on the most discriminative parts in the global context (*e.g.*, hats, shoes, and bags) and is vulnerable to viewpoint changes and occlusions. Thanks to PGLR, which refines global pseudo-labels with rich fine-grained information, our model learns discriminative information from each part and captures diverse regions.

### 4.3. Comparison with State-of-the-Arts

We compare our method with state-of-the-art unsupervised re-ID methods on Market-1501, MSMT17, and VeRi-776, and all the results are in Table 3.

We first compare our PPLR without the inter-camera contrastive loss (*i.e.*,  $\lambda_{cam} = 0$ ) with unsupervised methods that do not use camera labels: BUC [24], MMCL [47], HCT [61], MMT [8], SpCL [9], GCL [3], and RLCC [64]. SpCL employs a robust clustering criterion by identifying unreliable clusters to simulate accurate pseudo labels.

Method		Market-1501				MSMT17				VeRi-776			
		mAP	R1	R5	R10	mAP	R1	R5	R10	mAP	R1	R5	R10
<i>Unsupervised methods without any labels</i>													
BUC [24]	AAAI'19	38.3	66.2	79.6	84.5	-	-	-	-	-	-	-	-
MMCL [47]	CVPR'20	45.5	80.3	89.4	92.3	11.2	35.4	44.8	49.8	-	-	-	-
HCT [61]	CVPR'20	56.4	80.0	91.6	95.2	-	-	-	-	-	-	-	-
MMT [8]	ICLR'20	74.3	88.1	96.0	97.5	-	-	-	-	-	-	-	-
GCL [3]	CVPR'21	66.8	87.3	93.5	95.5	21.3	45.7	58.6	64.5	-	-	-	-
SpCL [9]	NeurIPS'20	73.1	88.1	95.1	97.0	19.1	42.3	56.5	68.4	36.9	79.9	86.8	89.9
RLCC [64]	CVPR'21	77.7	90.8	96.3	97.5	27.9	56.5	68.4	73.1	39.6	83.4	88.8	90.9
<b>PPLR (Ours)</b>	This work	<b>81.5</b>	<b>92.8</b>	<b>97.1</b>	<b>98.1</b>	<b>31.4</b>	<b>61.1</b>	<b>73.4</b>	<b>77.8</b>	<b>41.6</b>	<b>85.6</b>	<b>91.1</b>	<b>93.4</b>
<i>Unsupervised methods using camera labels</i>													
SSL [25]	CVPR'19	37.8	71.7	83.8	87.4	-	-	-	-	-	-	-	-
JVTC [19]	ECCV'20	47.5	79.5	89.2	91.9	17.3	43.1	53.8	59.4	-	-	-	-
IICS [55]	CVPR'21	72.9	89.5	95.2	97.0	26.9	56.4	68.8	73.4	-	-	-	-
CAP [50]	AAAI'21	79.2	91.4	96.3	97.7	36.9	67.4	78.0	81.4	-	-	-	-
ICE [2]	ICCV'21	82.3	93.8	97.6	98.4	38.9	70.2	80.5	84.4	-	-	-	-
<b>PPLR (Ours)</b>	This work	<b>84.4</b>	<b>94.3</b>	<b>97.8</b>	<b>98.6</b>	<b>42.2</b>	<b>73.3</b>	<b>83.5</b>	<b>86.5</b>	<b>43.5</b>	<b>88.3</b>	<b>92.7</b>	<b>94.4</b>

Table 3. Comparison with the state-of-the-art methods on Market1501, MSMT17, and VeRi-776.

MMT and RLCC refine noisy pseudo-labels using the predictions of an auxiliary network and a cluster consensus matrix, respectively. However, all these methods consider only global features and neglect the fine-grained information essential to person re-ID. On the other hand, our method considers the complementary relationship between global and part features and collectively mitigates the label noise for global features through the ensemble of the part predictions. As shown in Table. 3, our method surpasses prior state-of-the-art methods on all the benchmarks with +3.8%, +3.5%, and +2.0% of mAP than RLCC on Market-1501, MSMT17, and VeRi-776, respectively.

We also compare our PPLR with unsupervised methods trained with camera labels: SSL [25], JVTC [19], IICS [55], CAP [50], and ICE [2]. IICS and CAP exploit intra- and inter-camera similarities and train the model based on each similarity to reduce large intra-class variance from different camera views. ICE leverages inter-instance pairwise similarities based on data augmentation strategies to boost contrastive learning schemes. Our method also utilizes various feature similarity structures, but we focus on exploiting both the global and local similarity information desirable for fine-grained person re-ID. Additionally, unlike the other methods that directly use pseudo-labels computed by feature similarities, our method considers the complementary information between the feature similarity structures. As shown in Table. 3, our method significantly outperforms state-of-the-art methods by a remarkable margin; PPLR scores higher in mAP than ICE by +2.1% and +3.3% on Market-1501 and MSMT17, respectively.

## 5. Discussion

While we have shown the effectiveness of our method, it still has a limitation to overcome. Like other part-based methods [7, 42, 67], we extract part features by partitioning a feature map uniformly. Since not all images are aligned to a human body, however, part features in the same part index

may not represent the same body parts. Even though they capture discriminative information, misaligned part features may output ranked lists with noisy information and produce low cross agreement scores. To overcome this, a promising solution would be using semantic matching or human parsing techniques to construct feature spaces that represent a shared, similar semantic part of a person.

In addition, the re-ID techniques may potentially bring negative impacts, such as infringement of privacy due to abuse of a surveillance system. Related researchers and users should be attentive to using the technology in an appropriate manner while considering ethical issues. DukeMTMC-reID [33] has been taken down due to ethical concerns and should no longer be used.

## 6. Conclusion

In this paper, we have proposed a Part-based Pseudo Label Refinement (PPLR) framework for unsupervised person re-identification. Our method exploits both the global and local context of an image and alleviates the label noise for each feature space using the complementary relationship between the global and part features. We have introduced a cross agreement score to leverage reliable complementary information, and based on this, we have proposed agreement-aware label smoothing and part-guided label refinement. We have conducted extensive ablation studies, including qualitative analysis, to validate the effectiveness of the proposed method. Furthermore, extensive experiments have shown that our framework outperforms prior state-of-the-art methods on several benchmarks.

**Acknowledgement** This work was supported in part by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. IITP-2015-0-00199, Proximity computing and its applications to autonomous vehicle, image search, and 3D printing) and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2021R1C1C1012540).

## References

- [1] Haw-Shiuan Chang, Erik Learned-Miller, and Andrew McCallum. Active bias: Training more accurate neural networks by emphasizing high variance samples. In *NeurIPS*, 2017. 2
- [2] Hao Chen, Benoit Lagadec, and Francois Bremond. Ice: Inter-instance contrastive encoding for unsupervised person re-identification. In *ICCV*, 2021. 1, 2, 4, 8
- [3] Hao Chen, Yaohui Wang, Benoit Lagadec, Antitzia Dantcheva, and Francois Bremond. Joint generative and contrastive learning for unsupervised person re-identification. In *CVPR*, 2021. 7, 8
- [4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 6
- [5] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *CVPR*, 2018. 2
- [6] 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. 3, 6, 13
- [7] Yang Fu, Yunchao Wei, Guanshuo Wang, Yuqian Zhou, Honghui Shi, and Thomas S Huang. Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In *ICCV*, 2019. 1, 2, 8
- [8] 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, 2, 5, 7, 8, 12
- [9] 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 *NeurIPS*, 2020. 1, 2, 7, 8
- [10] Aritra Ghosh, Himanshu Kumar, and PS Sastry. Robust loss functions under label noise for deep neural networks. In *AAAI*, 2017. 2
- [11] Jianyuan Guo, Yuhui Yuan, Lang Huang, Chao Zhang, Jin-Ge Yao, and Kai Han. Beyond human parts: Dual part-aligned representations for person re-identification. In *ICCV*, 2019. 2
- [12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6
- [13] Dan Hendrycks, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel. Using trusted data to train deep networks on labels corrupted by severe noise. In *NeurIPS*, 2018. 2
- [14] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 4
- [15] Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, Teddy Furon, and Ondrej Chum. Efficient diffusion on region manifolds: Recovering small objects with compact cnn representations. In *CVPR*, 2017. 4
- [16] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *NeurIPS*, 2015. 2
- [17] Mahdi M Kalayeh, Emrah Basaran, Muhittin Gökm̄en, Mustafa E Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In *CVPR*, 2018. 2
- [18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6
- [19] Jianing Li and Shiliang Zhang. Joint visual and temporal consistency for unsupervised domain adaptive person re-identification. In *ECCV*, 2020. 8
- [20] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deep-reid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014. 1
- [21] Wei Li, Xiatian Zhu, and Shaogang Gong. Person re-identification by deep joint learning of multi-loss classification. In *IJCAI*, 2017. 1
- [22] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *CVPR*, 2018. 1, 2
- [23] Shan Lin, Haoliang Li, Chang-Tsun Li, and Alex Chichung Kot. Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification. In *BMVC*, 2018. 2
- [24] Yutian Lin, Xuanyi Dong, Liang Zheng, Yan Yan, and Yi Yang. A bottom-up clustering approach to unsupervised person re-identification. In *AAAI*, 2019. 1, 2, 3, 7, 8
- [25] Yutian Lin, Lingxi Xie, Yu Wu, Chenggang Yan, and Qi Tian. Unsupervised person re-identification via softened similarity learning. In *CVPR*, 2020. 1, 2, 8
- [26] Xinchen Liu, Wu Liu, Tao Mei, and Huadong Ma. A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In *ECCV*, 2016. 5
- [27] Michal Lukasik, Srinadh Bhojanapalli, Aditya Menon, and Sanjiv Kumar. Does label smoothing mitigate label noise? In *ICML*, 2020. 5
- [28] Hao Luo, Wei Jiang, Youzhi Gu, Fuxu Liu, Xingyu Liao, Shenqi Lai, and Jianyang Gu. A strong baseline and batch normalization neck for deep person re-identification. *IEEE Transactions on Multimedia (TMM)*, 2019. 6
- [29] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. Pose-guided feature alignment for occluded person re-identification. In *ICCV*, 2019. 2
- [30] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *CVPR*, 2017. 2
- [31] Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. In *ICLR Workshops*, 2017. 5
- [32] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *ICML*, 2018. 2
- [33] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV Workshops*, 2016. 8
- [34] M Saquib Sarfraz, Arne Schumann, Andreas Eberle, and Rainer Stiefelhagen. A pose-sensitive embedding for per-

- son re-identification with expanded cross neighborhood re-ranking. In *CVPR*, 2018. 4
- [35] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017. 7
- [36] Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. In *NeurIPS*, 2019. 2
- [37] Jianlou Si, Honggang Zhang, Chun-Guang Li, Jason Kuen, Xiangfei Kong, Alex C Kot, and Gang Wang. Dual attention matching network for context-aware feature sequence based person re-identification. In *CVPR*, 2018. 2
- [38] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *arXiv preprint arXiv:2007.08199*, 2020. 2
- [39] Liangchen Song, Cheng Wang, Lefei Zhang, Bo Du, Qian Zhang, Chang Huang, and Xinggang Wang. Unsupervised domain adaptive re-identification: Theory and practice. *Pattern Recognition*, 2020. 2
- [40] Chi Su, Jiani Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Pose-driven deep convolutional model for person re-identification. In *ICCV*, 2017. 2
- [41] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoungh Mu Lee. Part-aligned bilinear representations for person re-identification. In *ECCV*, 2018. 2
- [42] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, 2018. 1, 3, 8
- [43] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016. 5
- [44] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, 2017. 6
- [45] Arash Vahdat. Toward robustness against label noise in training deep discriminative neural networks. In *NeurIPS*, 2017. 2
- [46] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research (JMLR)*, 2008. 4, 7
- [47] Dongkai Wang and Shiliang Zhang. Unsupervised person re-identification via multi-label classification. In *CVPR*, 2020. 1, 2, 7, 8
- [48] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *ACM MM*, 2018. 2, 3
- [49] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *CVPR*, 2018. 2
- [50] 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, 6, 8, 12
- [51] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *ICCV*, 2019. 2
- [52] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, 2018. 2, 5
- [53] Longhui Wei, Shiliang Zhang, Hantao Yao, Wen Gao, and Qi Tian. Glad: Global-local-alignment descriptor for pedestrian retrieval. In *ACM MM*, 2017. 1, 2
- [54] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *CVPR*, 2015. 2
- [55] Shiyu Xuan and Shiliang Zhang. Intra-inter camera similarity for unsupervised person re-identification. In *CVPR*, 2021. 4, 8
- [56] Fan Yang, Ryota Hinami, Yusuke Matsui, Steven Ly, and Shin'ichi Satoh. Efficient image retrieval via decoupling diffusion into online and offline processing. In *AAAI*, 2019. 4
- [57] Fan Yang, Ke Yan, Shijian Lu, Huizhu Jia, Xiaodong Xie, and Wen Gao. Attention driven person re-identification. *Pattern Recognition*, 2019. 2
- [58] Qize Yang, Hong-Xing Yu, Ancong Wu, and Wei-Shi Zheng. Patch-based discriminative feature learning for unsupervised person re-identification. In *CVPR*, 2019. 2
- [59] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE TPAMI*, 2021. 1
- [60] Hong-Xing Yu, Wei-Shi Zheng, Ancong Wu, Xiaowei Guo, Shaogang Gong, and Jian-Huang Lai. Unsupervised person re-identification by soft multilabel learning. In *CVPR*, 2019. 1, 2
- [61] Kaiwei Zeng, Munan Ning, Yaohua Wang, and Yang Guo. Hierarchical clustering with hard-batch triplet loss for person re-identification. In *CVPR*, 2020. 2, 3, 7, 8
- [62] Yunpeng Zhai, Shijian Lu, Qixiang Ye, Xuebo Shan, Jie Chen, Rongrong Ji, and Yonghong Tian. Ad-cluster: Augmented discriminative clustering for domain adaptive person re-identification. In *CVPR*, 2020. 1
- [63] 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, 2, 5, 12
- [64] Xiao Zhang, Yixiao Ge, Yu Qiao, and Hongsheng Li. Refining pseudo labels with clustering consensus over generations for unsupervised object re-identification. In *CVPR*, 2021. 1, 2, 3, 7, 8
- [65] Zhilu Zhang and Mert R Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *NeurIPS*, 2018. 2
- [66] Liming Zhao, Xi Li, Yueling Zhuang, and Jingdong Wang. Deeply-learned part-aligned representations for person re-identification. In *ICCV*, 2017. 2
- [67] Feng Zheng, Cheng Deng, Xing Sun, Xinyang Jiang, Xiaowei Guo, Zongqiao Yu, Feiyue Huang, and Rongrong Ji. Pyramidal person re-identification via multi-loss dynamic training. In *CVPR*, 2019. 1, 2, 3, 8

- [68] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jing-dong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015. 5
- [69] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016. 1
- [70] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In *CVPR*, 2019. 1
- [71] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *CVPR*, 2017. 4, 6, 13
- [72] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *AAAI*, 2020. 6
- [73] Zhun Zhong, Liang Zheng, Shaozi Li, and Yi Yang. Generalizing a person retrieval model hetero- and homogeneously. In *ECCV*, 2018. 2
- [74] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *CVPR*, 2019. 2

## Appendix

### 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. 7. Larger values of  $N_p$  reduce the receptive field of the part feature to contain limited clues 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 simply 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 cross agreement scores, agreement-aware label smoothing adjusts the label distribution while considering the context of each part, leading to proper part feature learning.

#### 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<sup>2</sup> [63] and MMT<sup>3</sup> [8]. We analyze the number of training parameters, training stage time, and clustering stage time of each method, and the results are in Tab. 4. 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.

#### A.3. Qualitative Results

To further analyze cross agreement scores, we visualize images that have low- and top-50 cross agreement scores on Market-1501. As shown in Fig. 8, 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

<sup>2</sup><https://github.com/YunpengZhai/MEB-Net>

<sup>3</sup><https://github.com/yxgeee/MMT>

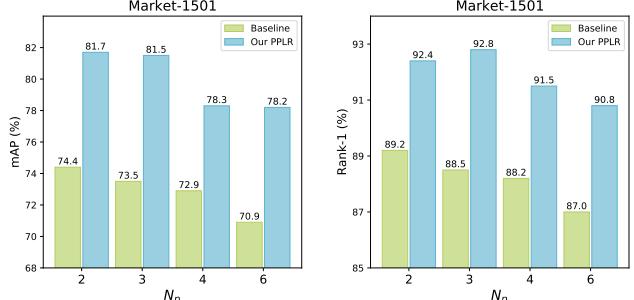


Figure 7. Parameter analysis of  $N_p$  on Market-1501. ‘Baseline’ is the part-based unsupervised re-ID framework (Sec. 3.1).

Method	Parameters (M)	Clustering stage time (sec / epoch)	Training stage time (sec / iter)
MEB-Net [63]	56.867	120.731	0.998
MMT [8]	50.096	48.545	0.511
Baseline	29.668	36.103	0.194
PPLR	29.668	38.484	0.202

Table 4. 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.

work. Some misaligned parts with 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.

#### B. Camera-aware Proxy Details

When the camera labels are available, PPLR can optionally leverage the inter-camera contrastive loss with camera-aware proxies [50]. 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:

$$\mathbf{c}_{(a,b)} = \frac{1}{|S_{(a,b)}|} \sum_{i \in S_{(a,b)}} \mathbf{f}_i, \quad (12)$$

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.

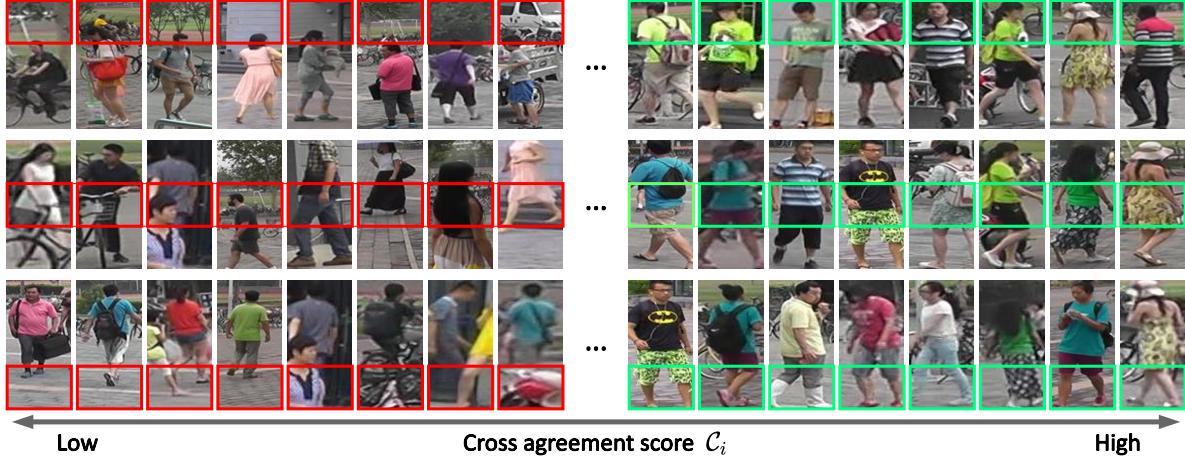


Figure 8. Visualization of images with low-50 and top-50 cross agreement scores on Market-1501. Very similar and duplicated images were excluded to show various cases.

## 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 [71] for clustering, where  $k$  is set to 30. For parameters of DBSCAN [6], 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 VeVi-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.

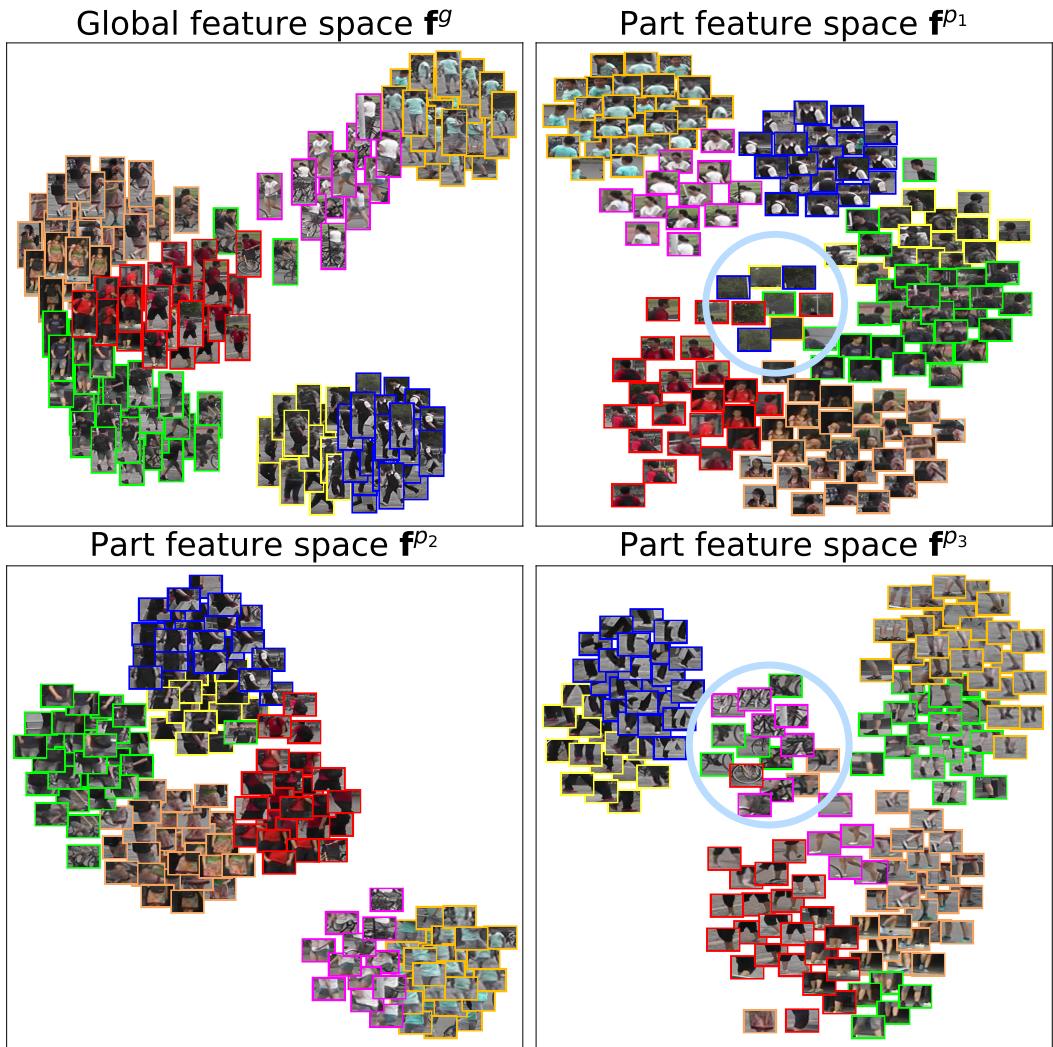


Figure 9. The large version of Fig. 2.

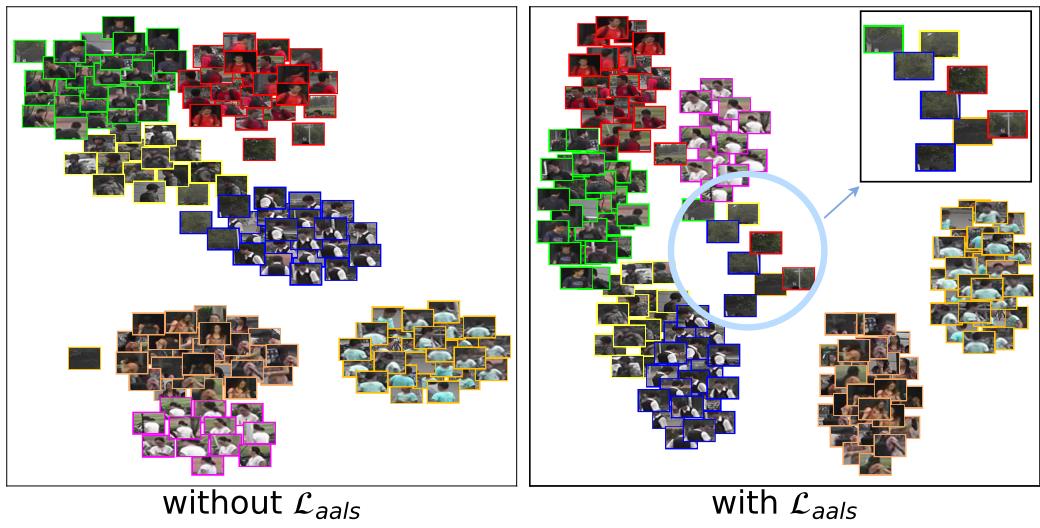


Figure 10. The large version of Fig. 5.