

Towards Precise Intra-camera Supervised Person Re-Identification

Menglin Wang¹, Baisheng Lai², Haokun Chen¹, Jianqiang Huang², Xiaojin Gong*^{†1}, and Xian-Sheng Hua²

¹Zhejiang University, China, ²Damo Academy, Alibaba Group

lynnwang6875@gmail.com; baisheng.lbs@alibaba-inc.com; chenhaokun@zju.edu.cn;
jianqiang.hjq@alibaba-inc.com; gongxj@zju.edu.cn; huaxiansheng@gmail.com;

Abstract

Intra-camera supervision (ICS) for person re-identification (Re-ID) assumes that identity labels are independently annotated within each camera view and no inter-camera identity association is labeled. It is a new setting proposed recently to reduce the burden of annotation while expect to maintain desirable Re-ID performance. However, the lack of inter-camera labels makes the ICS Re-ID problem much more challenging than the fully supervised counterpart. By investigating the characteristics of ICS, this paper proposes jointly learned camera-specific non-parametric classifiers, together with a hybrid mining quintuplet loss, to perform intra-camera learning. Then, an inter-camera learning module consisting of a graph-based ID association step and a Re-ID model updating step is conducted. Extensive experiments on three large-scale Re-ID datasets show that our approach outperforms all existing ICS works by a great margin. Our approach performs even comparable to state-of-the-art fully supervised methods in two of the datasets.

1. Introduction

Person re-identification (Re-ID) is the task of matching images of the same person across disjoint cameras. Because of its significance in surveillance, this task has attracted broad research interest in recent years. Most previous works focus on fully supervised [10, 19, 28, 1, 13] and unsupervised [4, 2, 32, 6, 22, 26] settings. The performance of supervised person Re-ID has been greatly improved due to the development of deep learning techniques. However, these methods need a large amount of full annotations that are expensive and time-consuming to obtain, leading to poor

*The corresponding author.

[†]This work was supported by Major Scientific Research Project of Zhejiang Lab (No. 2019DB0ZX01).

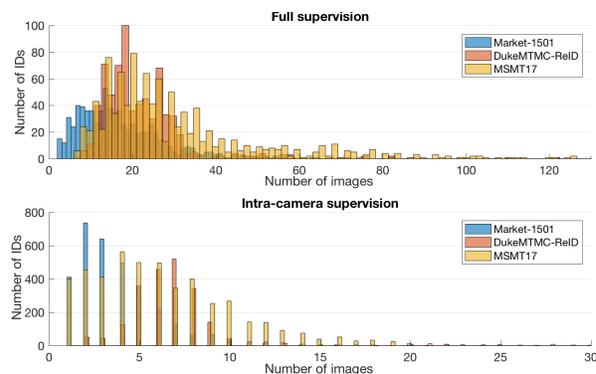


Figure 1. Comparison of the ID histograms under full and intra-camera supervisions. The histograms are computed in three large-scale datasets: Market-1501 [29], DukeMTMC-reID [17, 31], and MSMT17 [21].

scalability in real-world deployments. Conversely, unsupervised methods require no annotations but their performance is still far from satisfactory.

This paper aims to learn a person Re-ID model under intra-camera supervision (ICS), which is a new supervised setting proposed very recently [36, 16]. It assumes that identity labels are independently annotated within each camera and no inter-camera identity association is labeled. Since the ID association across cameras is known as the most time-consuming step for manual annotation, ICS can greatly reduce annotation costs and make the Re-ID techniques more scalable. Nevertheless, the per-camera independent labeling nature brings up two challenges: 1) Although IDs within each camera are labeled, most IDs in ICS contain much less samples than those in full supervision, as shown in Figure 1. 2) The lack of inter-camera labels result in more difficulties when dealing with appearance variations in different cameras.

Therefore, how to properly exploit the labeled data within cameras and mine the unlabeled relations across

cameras are two crucial problems in ICS. Existing works [36, 35, 16, 15] address these problems, respectively, from intra- and inter-camera learning perspectives. For intra-camera learning, MTML [36] and MATE [35] construct a multi-branch network that consists of a shared feature extraction backbone and multiple classification branches. Each branch learns a parametric softmax classifier for a specific camera view. However, the parametric classifiers could become ineffective when IDs contain only a couple of samples, which is a situation commonly occurred in ICS. Alternatively, PCSL [16] and ACAN [15] perform the intra-camera learning by adopting the widely used triplet loss [8], which mines the hardest positive and negative pairs within each training batch. The triplet loss takes only one negative sample each time and does not interact with other negative classes, leading to inferior performance and slow convergence [18]. These ineffective intra-camera learning consequently hurts the inter-camera learning part, leading to a Re-ID performance much lower than the counterparts in full supervision.

This paper aims to design a precise model for ICS person Re-ID. Considering that each camera has a variable number of IDs and most IDs have only few samples, we propose to jointly learn a set of camera-specific non-parametric classifiers for the intra-camera learning. These non-parametric classifiers, which are implemented with a shared feature extraction backbone and an external memory bank, are independent to the number of identity classes, enabling our network to entirely focus on the learning of discriminative features. Based on the memory bank, we also propose a quintuplet loss, which takes both in-batch samples and memory-stored ID centroids into account, to boost the intra-camera learning performance. These strategies help us to exploit the per-camera labeled data thoroughly and achieve a Re-ID model with considerable discrimination ability. Afterwards, an inter-camera learning step is followed that aims to improve the Re-ID model by mining ID relationships across cameras. To this end, we design a graph-based strategy for ID association and pseudo labeling. The pseudo labels are further utilized as ground-truth to train our network in a fully supervised manner and gain a better Re-ID model.

Although we address the ICS problem from intra- and inter-camera learning perspectives as existing ICS works [15, 16, 36, 35], our approach is distinct in the learning strategies and achieves much higher Re-ID performance. Specifically, our contributions are as follow:

- 1) We propose jointly learned camera-specific non-parametric classifiers and a quintuplet loss for intra-camera learning. These designs are customized not only for the characteristics of ICS but also for the memory assisted network architecture. They enable our intra-camera learning module to achieve a Re-ID performance better than existing ICS full models [15, 16, 36, 35] that consider both intra-

and inter-camera learning parts.

- 2) Benefited from the high discriminative model obtained above, together with the proposed graph-based association strategy, we are able to get desirable pseudo labels in inter-camera learning. These labels enable us to effectively train our Re-ID model, which is built on the BNNeck augmented network architecture [13], in a fully supervised manner. Riding the wave of architectures successfully applied in the supervised Re-ID task also boosts the performance further.

- 3) Extensive experiments on three large-scale Re-ID datasets including Market-1501 [29], DukeMTMC-reID [17, 31], and MSMT17 [21], show that the proposed approach outperforms previous ICS works by a great margin. Our performance is even comparable to fully supervised methods on the first two datasets.

2. Related Work

2.1. Person Re-identification

Fully supervised person Re-ID has made significant progress relying on the success of deep learning techniques. However, it remains to be an unsolved problem due to challenges arising from cluttered background, occlusion, as well as variations in illumination, pose, and viewpoint. Recent methods have exploited part-based features [19], human semantics [28], attention mechanisms [1], or data generation [30] to tackle the challenges. These methods often lead to complex network architectures. An exceptional work is Bag of Tricks (BoT) [13] that achieves the state-of-the-art performance by augmenting the baseline network with a simple BNNeck component. In this paper, we construct our network upon the BoT model, i.e. the BNNeck augmented ResNet-50 [7], to keep our backbone simple yet effective.

Unsupervised person Re-ID has attracted a lot of research interest in recent years. Existing methods can be roughly categorized into two groups. One is based on domain adaptation techniques [4, 32, 5] that transfer knowledge from labeled source domain to unlabeled target domain. The other group [6, 12, 22] is purely unsupervised that requires no external labeled data. Both types of the methods commonly perform a step to associate IDs across cameras via K-nearest neighbors [32], clustering [6, 12], or graph [22] based strategies. Our ICS work also adopts a graph-based ID association step. But, in contrast to use a graph-weighted loss [22], we formulate the association as a problem of finding connected components in a sparse graph.

Semi-supervised person Re-ID aims to learn a Re-ID model from both labeled and unlabeled data [25]. Intra-camera supervision (ICS) is a special semi-supervised setting proposed recently. All existing ICS works address the problem by considering both intra- and inter-camera learning perspectives. For supervised intra-camera learning,

PCSL [16] and ACAN [15] take the extensively used triplet loss [8], while MTML [36] and MATE [35] design a multi-branch structure to learn parametric classifiers. For inter-camera learning, ACAN [15] develops a multi-camera adversarial learning approach to reduce the cross-camera data distribution discrepancy, PCSL [16] utilizes a soft-labeling scheme, MTML [36] and MATE [35] adopt multi-label learning strategies. Although we address the ICS problem from intra- and inter-camera learning perspectives as well, we propose distinct learning strategies in both parts and achieve much higher Re-ID performance.

Difference from memory-assisted Re-ID methods.

Although memory assisted non-parametric classification has been adopted in unsupervised Re-ID [32] and semi-supervised Re-ID tasks [25], our approach is designed specific to the intra-camera supervised setting. Specifically, in contrast to [32, 25] that use a memory to store instance features and apply one classifier for all images, we construct multiple non-parametric classifiers, each of which corresponds to one camera and all of which are learned jointly. In addition, memory is also used in PCSL [16], but PCSL uses ID centroids stored in the memory to weight a parametric classifier for cross-camera soft-labeling, while we use non-parametric classifiers for intra-camera learning.

2.2. Parametric and Non-parametric Classifiers

Parametric classifiers in this work refer to those implemented by fully connected (FC) layers in a deep neural network (DNN), usually trained with a cross-entropy softmax classification loss. Such classifiers have been successfully used in fully supervised person Re-ID [19, 27, 13]. However, they have the following drawbacks: 1) The training process pays much attention to learning parameters in the FC layers that are abandoned during inference for person Re-ID [27], making the learned feature representation less discriminative for test data. 2) The classifiers can not be learned effectively when there are a large number of identities while each identity only has a small number of instances [24].

Non-parametric classifiers in a DNN are commonly implemented via a non-parametric variant of the softmax function. It is first proposed in a fully supervised person search task [24] and extensively adopted in unsupervised [23, 32, 33] and semi-supervised [3, 25] learning. A common challenge in these tasks is that the number of classes is huge but each class contains only one or few samples. A DNN equipped with the non-parametric classifiers makes its parameters independent to the class number so that the training process entirely focus on the feature representation learning. This property benefits the Re-ID performance since a person is re-identified only based on the distance of features at the test time. Nevertheless, the non-parametric model may overfit more easily when training

data is abundant enough [24]. Considering that each camera has a variable number of IDs and most IDs contain only a few of samples due to the per-camera labeling manner in ICS, we adopt non-parametric classifiers for our intra-camera learning. While in inter-camera learning where IDs contain more samples after association, we choose the parametric variant.

3. The Proposed Method

The intra-camera supervision assumes that identity labels are independently annotated within each camera view and no inter-camera identity association is provided. Suppose there are C cameras in a dataset. We denote the set of the c -th camera by $\mathcal{D}_c = \{(x_i, y_i, c_i)\}_{i=1}^{M_c}$, in which image x_i is annotated with an identity label $y_i \in \{1, \dots, N_c\}$ and a camera label $c_i = c \in \{1, \dots, C\}$. M_c and N_c are, respectively, the number of total images and IDs in this camera view. $N = \sum_{c=1}^C N_c$ is the total ID number directly accumulated over all cameras. It should be noted that the identities in different cameras are partially overlapped. That is, a same person may appear in two or more camera views, but it could be assigned with different IDs due to the per-camera independent labeling manner. Given this training set $\mathcal{D} = \bigcup_{c=1}^C \mathcal{D}_c$, we aim to learn a Re-ID model that can well discriminate both intra- and inter-camera identities.

To this end, we develop our method from both intra- and inter-camera learning perspectives. The overall framework is shown in Figure 2. An image is first fed into a backbone network for feature extraction. The extracted feature goes through an additional feature embedding layer and then classified by camera-specific non-parametric classifiers that are implemented via an external memory bank, together with an ID classification loss and a quintuplet loss, for intra-camera learning. The memory bank stores the centroid feature of each ID, which is of moderate discrimination ability after intra-camera learning. Then, the ID centroids are used for ID association and pseudo labeling across cameras. In inter-camera learning, the same backbone is adopted to extract features, along with a classifier parameterized by a FC layer to classify images into their pseudo identity classes.

3.1. Intra-camera Learning

When considering the Re-ID problem within an individual camera, it can be treated as a fully supervised classification task. Therefore, it is reasonable to formulate the intra-camera learning as a fully supervised multi-task classification problem that learns multiple per-camera ID classifiers jointly. However, the number of IDs in each camera varies a lot and most IDs have only several samples, making parametric classifiers implemented via multi-branch fully-connected architectures [36, 35] ineffective. Thus, we opt to use non-parametric variants in this work.

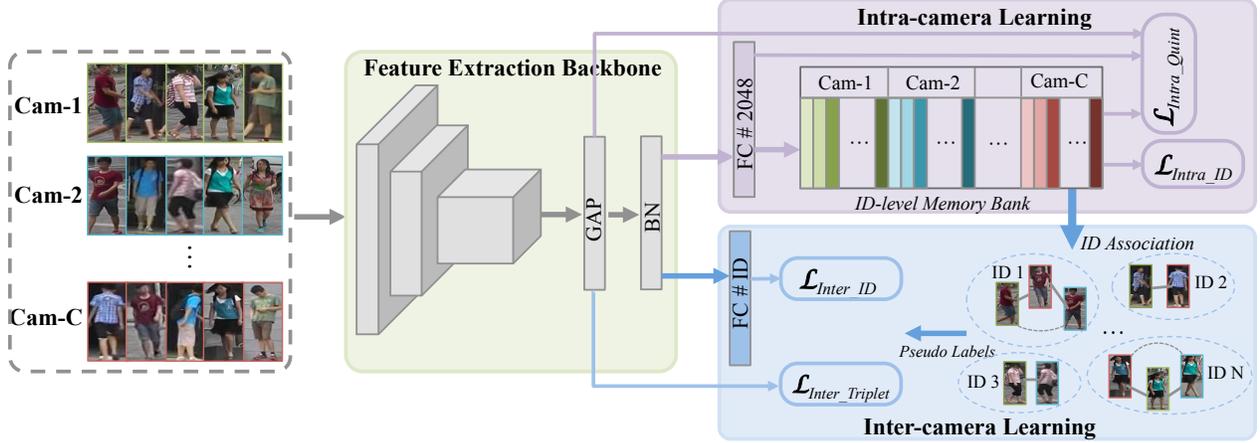


Figure 2. An overview framework of the proposed method. It consists of a feature extraction backbone, together with an intra-camera learning branch and an inter-camera learning branch. In intra-camera learning, features are classified via camera-specific non-parametric classifiers implemented via an external memory bank, optimized by \mathcal{L}_{Intra_ID} and $\mathcal{L}_{Intra_Quint}$. In inter-camera learning, IDs are associated across cameras and pseudo labeled, which are further used to update the Re-ID model by optimizing \mathcal{L}_{Inter_ID} and $\mathcal{L}_{Inter_Triplet}$.

We design a number of non-parametric classifiers, each of which performs the fully supervised ID classification within one camera view. All non-parametric classifiers share a feature extraction backbone and an external memory bank so that the classifiers are learned jointly. Each classifier, termed as a *camera-specific non-parametric classifier*, optimizes an ID classification loss that pulls an image close to the centroid of its labeled ID while pushes away from the centroids of all other IDs in the same camera. By considering the distances between samples and centroids, the ID loss is good at separating a majority of samples, but may still fail for hard ones. Regarding to this reason, we also propose a hybrid mining quintuplet loss to improve the inter-class separability. The details are introduced in the followings.

3.1.1 Camera-specific Non-parametric Classifiers

As illustrated in Figure 2, our network consists of a feature extraction backbone, a FC embedding layer, together with an external memory bank. Based upon this network structure, we design camera-specific non-parametric classifiers to perform the classification tasks within each camera view.

Formally, when an image is input, the FC embedding layer outputs a d -dimensional feature vector ($d = 2048$ in our work). The memory bank $\mathcal{K} \in R^{d \times N}$ stores the up-to-date features of all accumulated IDs and each column corresponds to one ID. During back-propagation, the memory bank is updated by

$$\mathcal{K}[j] \leftarrow \mu \mathcal{K}[j] + (1 - \mu) f(x_i), \quad (1)$$

where $\mathcal{K}[j]$ is the j -th column of the memory. $f(x_i)$ is a L_2 normalized feature extracted from image x_i that belongs to the j -th ID. $\mu \in [0, 1]$ is an updating rate. After each update,

$\mathcal{K}[j]$ is scaled to having unit L_2 norm. The updated feature in each column can be interpreted as the centroid of an identity class in the feature space, which is a d -dimensional unit hypersphere.

Given the image x_i , together with its annotated intra-camera identity label y_i and camera label c_i . The corresponding global ID index j is obtained by $j = A + y_i$, where $A = \sum_{k=1}^{c_i-1} N_k$ is the total ID number accumulated from the first to the $c_i - 1$ -th camera view. Then, the probability of classifying x_i into the j -th ID is defined by a non-parametric softmax function

$$p(j|x_i) = \frac{\exp(\mathcal{K}[j]^T f(x_i)/\tau)}{\sum_{k=A+1}^{A+N_{c_i}} \exp(\mathcal{K}[k]^T f(x_i)/\tau)}, \quad (2)$$

where τ is the temperature controlling the smoothness of probability distribution.

Note that the non-parametric classifier defined above is camera-specific, because the sum in the denominator is over the IDs within the same camera view only. In contrast to existing works [24, 23, 32, 5] that considers all entries in their memory bank, our formulation only takes those belonging to the same camera into account while ignores the IDs in all other cameras. Thus, each non-parametric classifier is responsible for the classification task in a specific camera.

To jointly learn all the camera-specific identity classification tasks together, the objective, termed as the *intra-camera ID loss*, is formulated to minimize the negative log-likelihood of training images in all cameras. That is,

$$\mathcal{L}_{Intra_ID} = - \sum_{c=1}^C \left(\frac{1}{|\mathcal{D}_c|} \sum_{(x_i, y_i, c_i) \in \mathcal{D}_c} \log p(j|x_i) \right), \quad (3)$$

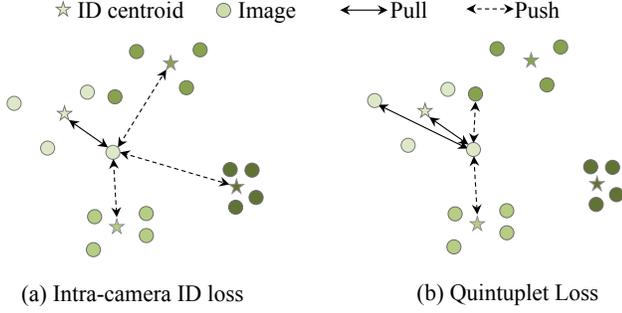


Figure 3. Illustration of intra-camera ID loss and quintuplet loss.

where $\frac{1}{|\mathcal{D}_c|}$ is a normalization coefficient placed to balance the variable number of images in different cameras.

3.1.2 A Hybrid Mining Quintuplet Loss

The intra-camera ID loss introduced above in essence pulls a sample close to its centroid and meanwhile pushes it away from the centroids of all other IDs within the same camera, as shown in Figure 3(a). This loss classifies most samples well, but may be failed for hard examples. To improve the inter-class separability and intra-class compactness further, we propose a quintuplet loss as a supplement. Our loss is constructed based upon the widely used triplet loss [8]. In addition to the triplet loss that mines hard positive and negative samples locally within each mini-batch, the quintuplet loss also incorporates the hard pairs stored in the global memory bank. It pulls an anchor close to both its centroid and the associated hard positive sample, and pushes away from both the hard negative sample and ID centroid, as illustrated in Figure 3(b).

Mathematically, in each mini-batch, we randomly select P identities and K instances of each identity, as the common practice [8]. For each anchor image x_a , we design a hybrid mining scheme that selects two instances and two identity centroids to form a quintuplet. The positive and negative instances are sampled to be the hardest ones within a mini-batch. In addition, we choose the positive ID centroid and the nearest negative ID centroid from the memory bank. Note that all instances and centroids are selected from the same camera as the anchor. Then, the intra-camera quintuplet loss is defined as follows:

$$\begin{aligned} \mathcal{L}_{Intra_Quint} = & \sum_{i=1}^P \sum_{a=1}^K \left[m_1 + \max_{p=1, \dots, K} \|g(x_a^i) - g(x_p^i)\| \right. \\ & \left. - \min_{n=1, \dots, K; j=1, \dots, P; j \neq i; c_a^i = c_n^j} \|g(x_a^i) - g(x_n^j)\| \right]_+ \\ & + \left[m_2 + \|f(x_a^i) - \mathcal{K}[A + y_a^i]\| \right. \\ & \left. - \min_{j=1, \dots, N_{c_a^i}; j \neq y_a^i} \|f(x_a^i) - \mathcal{K}[A + j]\| \right]_+, \end{aligned} \quad (4)$$

where m_1 and m_2 are two margins. $g(x_a)$, $g(x_p)$, and $g(x_n)$ are, respectively, the features of the anchor, positive and negative instances output from the global average pooling (GAP) layer in the backbone. $f(x_a)$ is the anchor's feature produced from the FC embedding layer as above-mentioned. Taking features from different layers is inspired by the BNNeck structure in [13] and shown to be effective in our experiments. In addition, $A + y_a$ is the global ID index in \mathcal{K} given the intra-camera ID y_a , $[\cdot]_+ = \max(0, \cdot)$, and $\|\cdot\|$ is the Euclidean distance.

3.1.3 The Loss for Intra-camera Learning

In summary, the loss for intra-camera learning is the sum of the intra-camera ID loss and the quintuplet loss:

$$\mathcal{L}_{Intra} = \mathcal{L}_{Intra_ID} + \mathcal{L}_{Intra_Quint}. \quad (5)$$

3.2. Inter-camera Learning

With the proposed camera-specific non-parametric ID classification loss and quintuplet loss, our Re-ID model achieves outstanding discrimination ability within cameras. Meanwhile, the model also gains certain capability in discriminating IDs across cameras through the joint learning of all per-camera classification tasks. However, due to the lack of explicit inter-camera correlations, the model is still weak at coping with variations in different cameras. To address this problem, we perform an inter-camera learning that consists of a cross-camera ID association step and a Re-ID model updating step.

3.2.1 Cross-camera ID Association

We formulate the cross-camera ID association task as a problem of finding connected components in a graph. We construct the graph based on two observations: 1) The more similar two IDs are, the more likely they are to be the same person. 2) Under the intra-camera supervised setting, each ID has no positive matches within the same camera, and at most one positive match in any associated camera. According to these observations, we construct an undirected graph $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ for association, where the vertex set \mathcal{V} denotes the IDs accumulated over all cameras and the edge set $E = \{e(i, j)\}$ indicates whether the i -th ID and the j -th ID is a positive pair or not. The edge $e(i, j)$ is defined by

$$e(i, j) = \begin{cases} 1, & \text{dist}(i, j) < T \wedge c(i) \neq c(j) \\ & \wedge i \in \mathcal{N}_1(j, c(i)) \wedge j \in \mathcal{N}_1(i, c(j)); \\ 0, & \text{otherwise.} \end{cases} \quad (6)$$

Here, $\text{dist}(i, j) = \|\mathcal{K}[i] - \mathcal{K}[j]\|$ is the Euclidean distance between two ID centroids stored in the memory bank, and $i, j \in \{1, \dots, N\}$. T is a threshold taken by ascendingly sorting the distances of ID pairs and choosing the S -th distance

value, implying to choose Top S similar pairs. $c(i)$ represents the camera that the i -th ID belongs to. $\mathcal{N}_1(j, c(i))$ is the 1-nearest neighbor of ID j in camera $c(i)$ and likewise for $\mathcal{N}_1(i, c(j))$. The last two conditions require i and j to be the reciprocal nearest neighbor of each other.

The constructed graph is sparsely connected. Therefore, we adopt a union-find algorithm [14] to find all connected components in the graph. All IDs within each component are associated and assigned with a same pseudo label. The pseudo labels are further used to update the Re-ID model.

3.2.2 Re-ID Model Updating

Taking all images and their pseudo labels, we treat the Re-ID task as a fully supervised problem and adopt the simple yet effective architecture in BoT [13] to update the Re-ID model. Specifically, based upon the feature extraction backbone learned at our intra-camera learning stage, we append a FC layer to learn a parametric ID classifier, as shown in Figure 2. The model is trained with the extensively used cross-entropy loss with the label smoothing scheme, termed as an inter-camera ID loss $\mathcal{L}_{Inter.ID}$ in our work, together with a batch-hard triplet loss $\mathcal{L}_{Inter.Triplet}$ applied to the features output from the GAP layer. Therefore, the total loss for inter-camera learning is:

$$\mathcal{L}_{Inter} = \mathcal{L}_{Inter.ID} + \mathcal{L}_{Inter.Triplet}. \quad (7)$$

Note that the above-mentioned inter-camera ID loss is using a parametric Softmax. Analogous to the intra-camera learning, we can also implement it via the memory-assisted non-parametric Softmax. However, the non-parametric version is prone to overfit when training data become abundant after ID association, leading to an inferior performance.

4. Experiments

4.1. Experiment Setting

Datasets and Evaluation Metrics. We evaluate the proposed method on three large-scale datasets: Market-1501 [29], DukeMTMC-reID [17, 31], and MSMT17 [21]. The numbers of cameras, IDs, and images contained in each dataset are reported in Table 1. To simulate the ICS setting, we generate intra-camera identity labels based on the provided full annotations. Table 1 also lists the accumulated total identity number under intra-camera supervision ($\#ID_{ICS}$), the averaged image-per-person (IP) value, and the averaged image-per-camera-per-person (ICP) value. For performance evaluation, we adopt the Cumulative Matching Characteristic (CMC) and mean Average Precision (mAP), as the common practice.

Implementation details. We adopt the ImageNet [11]-pretrained ResNet-50 [7] as our feature extraction backbone. Following [13], we remove the last spatial downsampling operation in the backbone to increase the size of

Dataset	#Cam	#ID	#Img	#ID _{ICS}	IP	ICP
Market-1501	6	751	12,936	3,262	17.23	3.97
DukeMTMC-reID	8	702	16,522	2,196	23.54	7.52
MSMT17	15	1,041	32,621	4,821	31.34	6.77

Table 1. Dataset statistics. #Cam, #ID, #Img, and #ID_{ICS} are the number of cameras, IDs, images, and accumulated IDs under ICS, respectively. IP is the averaged image-per-person value, ICP is the averaged image-per-camera-per-person value.

feature maps, and add a batch normalization (BN) layer after GAP. During both intra- and inter-camera learning, images are resized to 256×128 . Random flipping, cropping, and erasing are performed as data augmentation. The initial learning rate is 0.00035 and divided by 10 after 40 and 70 epochs. We choose Adam [9] as the optimizer with weight decay 0.0005. Training batch size is 64, with randomly selected 16 IDs and 4 images for each ID. In intra-camera learning, images within each mini-batch are sampled according to the per-camera labels, and the total number of epochs is 50. Temperature τ in Equation (2) is set to 0.067 (i.e.1/15). The margins m_1 and m_2 in Equation (4) are both empirically set to 0.3. In inter-camera learning, images in each mini-batch are sampled according to the generated pseudo ID labels. The total number of epochs is 120. In graph construction, we select Top S similar pairs as candidates for edge linking, where S is empirically set to be the same number as #ID_{ICS} listed in Table 1. Each experiment runs 5 times and the averaged performance is reported.

4.2. Ablation Study

4.2.1 Effectiveness of The Intra-camera Learning Part

Effectiveness of The Camera-specific Non-parametric Classifiers. First, we are curious about how well the camera-specific non-parametric classifiers perform. Therefore, three model variants are investigated, including \mathcal{M}_1 : a multi-branch network structure [36], in which each branch uses a classifier parameterized by a FC layer and optimized with a cross-entropy ID loss, for intra-camera learning; \mathcal{M}_2 : a non-parametric classifier but not camera-specific, that is, any image can be classified into all accumulated ID classes; \mathcal{M}_3 : the proposed camera-specific non-parametric classifiers with the intra-camera ID loss only.

Table 2 presents the comparison results in terms of mAP(%) and Rank-1(%). From the results we observe that the camera-specific non-parametric classifiers (\mathcal{M}_3) outperform the camera-agnostic counterpart (\mathcal{M}_2) by a great margin, showing that the camera-specific scheme plays an important role. In addition, the camera-specific non-parametric classifiers (\mathcal{M}_3) perform consistently better than the parametric counterpart (\mathcal{M}_1) on all datasets. Especially, it improves the performance by a significant margin (mAP +14.2% and Rank-1 +9.3%) on Market1501 which has a smaller ICP value than the other two datasets, as reported in Table 1 and Fig. 1. It indicates that the non-parametric clas-

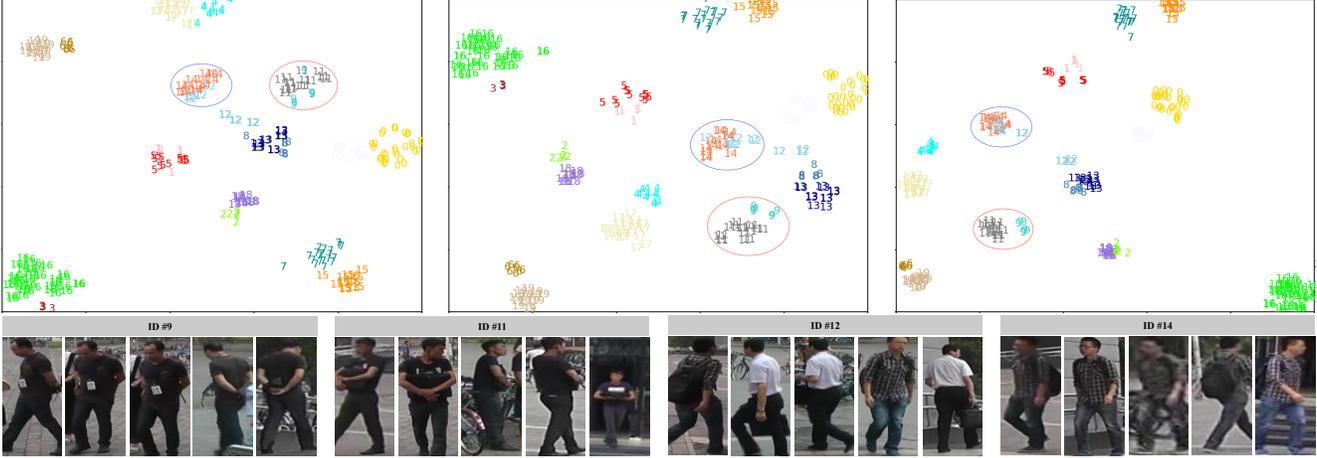


Figure 4. T-SNE visualization of the features learned by \mathcal{M}_3 , \mathcal{M}_5 , and \mathcal{M}_7 , which are respectively shown from left to right in the first row. Typical example images from IDs #9, #11, #12, #14 are presented in the bottom row.

sifiers are superior to the parametric counterpart especially when identity classes contain fewer examples.

Models	Market1501		DukeMTMC-ReID		MSMT17	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
\mathcal{M}_1	76.8	55.0	75.3	58.9	50.7	25.1
\mathcal{M}_2	45.5	30.7	33.3	29.1	10.4	6.0
\mathcal{M}_3	86.1	69.2	78.0	61.9	52.1	25.7
\mathcal{M}_4	86.8	71.9	79.1	64.0	54.3	28.1
\mathcal{M}_5	87.5	72.3	79.7	64.7	55.5	28.9
\mathcal{M}_6	91.5	79.7	80.6	65.5	49.7	23.4
\mathcal{M}_7	93.1	83.6	83.6	72.0	57.7	31.3
\mathcal{M}_8	94.1	85.9	87.4	77.2	75.3	52.2

Table 2. Comparison of the different model variants. \mathcal{M}_1 - \mathcal{M}_5 are intra-camera model variants, in which \mathcal{M}_1 adopts multi-branch parametric classifiers, \mathcal{M}_2 adopts camera-agnostic non-parametric classifiers, and \mathcal{M}_3 is our proposed camera-specific non-parametric classifiers. \mathcal{M}_4 is \mathcal{M}_3 with an additional triplet loss and \mathcal{M}_5 is \mathcal{M}_3 with the proposed quintuplet loss. \mathcal{M}_6 and \mathcal{M}_7 are, respectively, adopting a non-parametric and a parametric inter-camera learning part based upon \mathcal{M}_5 . \mathcal{M}_8 is a fully supervised version.

Effectiveness of The Quintuplet Loss. When validating the effectiveness of the proposed quintuplet loss, we investigate two model variants, which are \mathcal{M}_4 : the camera-specific non-parametric classifiers with the intra-camera ID loss and a batch-hard triplet loss [8]; and \mathcal{M}_5 : the camera-specific non-parametric classifiers with the intra-camera ID loss and the proposed quintuplet loss. From the results reported in Table 2, we observe that both \mathcal{M}_4 and \mathcal{M}_5 gain considerable improvements when compared to \mathcal{M}_3 that does not consider hard samples. Moreover, the model using the quintuplet loss (\mathcal{M}_5) performs consistently better than the one using the triplet loss (\mathcal{M}_4), demonstrating that it is effective to leverage information from both the local mini-batch and the global memory bank.

Market1501		DukeMTMC-ReID	
Precision	Recall	Precision	Recall
96.4%	75.9%	90.1%	74.3%

Table 3. Precision and recall of ID pairs associated by our inter-camera association strategy.

4.2.2 Effectiveness of The Inter-camera Learning Part

After the comparisons of all intra-camera learning components, we validate the effectiveness of the inter-camera learning part. Once getting pseudo labels, we have two options for inter-camera learning. They are, \mathcal{M}_6 that uses a non-parametric classifier as in intra-camera learning but is camera-agnostic, and \mathcal{M}_7 that adopts the parametric classifier as introduced in Section 3.2.2. The results in Table 2 demonstrate that the model \mathcal{M}_7 using the parametric classifier performs much better than the one \mathcal{M}_6 using the non-parametric variant. The reason we conjecture is that, the non-parametric version has much less parameters so that it is easier to get overfitting when most IDs contain abundant samples after ID association.

In addition, Table 2 also provides the results obtained by training the inter-camera learning branch in our network with entire ground-truth labels, which in essence is the fully supervised counterpart model \mathcal{M}_8 . This model indicates the upper bound performance that can be achieved by our Re-ID network architecture.

To investigate how the proposed graph-based ID association performs, we provide the precision and recall of the associated ID pairs using our ID association strategy. As shown in Table 3, the precision maintains a quite high value on both Market and Duke, showing the reliability of associated pairs for model learning. Besides, the recall is over 74%, indicating that most positive pairs are associated by our algorithm.

Methods	Reference	Market1501				DukeMTMC-ReID				MSMT17			
		R1	R5	R10	mAP	R1	R5	R10	mAP	R1	R5	R10	mAP
Fully supervised													
OSNet [34]	ICCV19	94.8	-	-	84.9	88.6	-	-	73.5	78.7	-	-	52.9
DGNet [30]	CVPR19	94.8	-	-	86.0	86.6	-	-	74.8	77.2	87.4	90.5	52.3
BoT [13]	CVPRW19	94.5	-	-	85.9	86.4	-	-	76.4	-	-	-	-
PCB [19]	ECCV18	93.8	-	-	81.6	83.3	-	-	69.2	68.2	-	-	40.4
Unsupervised													
ECN [32]	CVPR19	75.1	87.6	91.6	43.0	63.3	75.8	80.4	40.4	30.2	41.5	46.8	10.2
AE [5]	Arxiv19	81.6	91.9	94.6	58.0	67.9	79.2	83.6	46.7	32.3	44.4	50.1	11.7
BUC [12]	AAAI19	66.2	79.6	84.5	38.3	47.4	62.6	68.4	27.5	-	-	-	-
UGA [22]	ICCV19	87.2	-	-	70.3	75.0	-	-	53.3	49.5	-	-	21.7
Intra-camera supervised													
MTML [36]	ICCVW19	85.3	-	96.2	65.2	71.7	-	86.9	50.7	44.1	-	63.9	18.6
PCSL [16]	TCSVT20	87.0	94.8	96.6	69.4	71.7	84.7	88.2	53.5	48.3	62.8	68.6	20.7
ACAN [15]	Arxiv19	73.3	87.6	91.8	50.6	67.6	81.2	85.2	45.1	33.0	48.0	54.7	12.6
MATE [35]	Arxiv20	88.7	-	97.1	71.1	76.9	-	89.6	56.6	46.0	-	65.3	19.1
Precise-ICS	Ours	93.1	97.8	98.6	83.6	83.6	92.6	94.7	72.0	57.7	71.1	76.3	31.3

Table 4. Comparison with state-of-the-art methods.

4.2.3 Visualization of Learned Representations

We utilize t-SNE [20] to visualize the feature representations learned by our model components. Figure 4 presents the features of images in 20 IDs, respectively, learned by the camera-specific non-parametric classifiers (\mathcal{M}_3), the entire intra-camera learning model (\mathcal{M}_5), and the full model (\mathcal{M}_7). As Figure 4 shows, the camera-specific non-parametric classifiers (\mathcal{M}_3) gain considerable discrimination ability, but may mix up a number of difficult ID classes such as #9 and #11, #8 and #13, #1 and #5, as well as #12 and #14. With the additional quintuplet loss, the model \mathcal{M}_5 better separates these hard ID pairs that have very similar appearances, as sample images of #9 and #11 shown in the bottom row of Figure 4. The full model \mathcal{M}_7 , which incorporates both intra- and inter-camera learning, improves the intra-class compactness and inter-class separability further. The full model may still mix up a small number of IDs, but a part of the reason is because of the labeling noise, as the presented examples of #12 and #14.

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

In this section, we compare our approach (named as Precise-ICS) with all existing ICS person Re-ID methods, including MTML [36], PCSL [16], ACAN [15], and MATE [35]. From the results we observe that the proposed approach outperforms the other ICS methods by a great margin. More specifically, the mAP is 12.5%, 15.4%, and 10.6% higher and the Rank-1 accuracy is 4.4%, 6.7%, and 9.4% higher than the best performances obtained by other methods on Market1501, DukeMTMC-ReID, and MSMT17 respectively. Even the model using only the intra-camera learning part (Precise-ICS: \mathcal{M}_5 shown in Table 2) performs better than existing ICS methods, indicating that

the proposed components in our intra-camera learning can exploit per-camera labels much more thoroughly.

In addition, we also compare our work with state-of-the-art methods under different supervision settings, including four fully supervised methods and four unsupervised methods. As expected, our approach achieves much higher performance than the unsupervised methods no matter if they transfer knowledge from extra datasets or not. Meanwhile, our approach is better than an effective supervised method PCB [19], and is even comparable to recent supervised methods (DGNet [30], BoT [13]) on Market1501 and DukeMTMC-ReID. The results demonstrate the potential for the ICS Re-ID task to achieve high Re-ID performance while dramatically reduce labeling cost, making this supervision setting more scalable to real-world applications.

5. Conclusion

In this paper, we have proposed a new approach to address the person Re-ID problem under intra-camera supervision. The proposed network consists of a shared feature extraction backbone, together with two branches for intra- and inter-camera learning respectively. According to the per-camera labeling nature of ICS, we propose jointly learned camera-specific non-parametric classifiers and a hybrid mining quintuplet loss for intra-camera learning. The designed components exploit per-camera labels thoroughly so that our intra-camera learning part only performs better than most existing ICS methods. Benefited from the discrimination ability gained in this part, the inter-camera learning module can boost the Re-ID performance further by mining ID relationship across cameras. Our full model outperforms all ICS methods by a large margin, greatly reducing the gap to the fully supervised counterparts.

References

- [1] Tianlong Chen, Shaojin Ding, Jingyi Xie, Ye Yuan, Wuyang Chen, Yang Yang, Zhou Ren, and Zhangyang Wang. Abdnnet: Attentive but diverse person re-identification. In *ICCV*, 2019.
- [2] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. Deep association learning for unsupervised video person re-identification. In *BMVC*, 2018.
- [3] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. Semi-supervised deep learning with memory. In *ECCV*, 2018.
- [4] Weijian Deng, Liang Zheng, Qixiang Ye, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *CVPR*, 2018.
- [5] Yuhang Ding, Hehe Fan, Mingliang Xu, and Yi Yang. Adaptive exploration for unsupervised person re-identification. *arXiv preprint arXiv:1907.04194*, 2019.
- [6] Hehe Fan, Liang Zheng, Chenggang Yan, and Yi Yang. Unsupervised person re-identification: Clustering and fine-tuning. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 14(4):83, 2018.
- [7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.
- [8] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.
- [9] Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:2014*, 2014.
- [10] Elyor Kodirov, Tao Xiang, Zhenyong Fu, and Shaogang Gong. Person re-identification by unsupervised l_1 graph learning. In *European conference on computer vision*, pages 178–195. Springer, 2016.
- [11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- [12] Yutian Lin, Xuanyi Dong, Liang Zheng, Yan Yan, and Yi Yang. A bottom-up clustering approach to unsupervised person re-identification. In *AAAI*, 2019.
- [13] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *CVPRW*, 2019.
- [14] Md Mostofa Ali Patwary, Jean Blair, and Fredrik Manne. Experiments on union-find algorithms for the disjoint-set data structure. In *International Symposium on Experimental Algorithms*, pages 411–423. Springer, 2010.
- [15] Lei Qi, Lei Wang, Jing Huo, Yinghuan Shi, and Yang Gao. Adversarial camera alignment network for unsupervised cross-camera person re-identification. *arXiv preprint arXiv:1908.00862*, 2019.
- [16] Lei Qi, Lei Wang, Jing Huo, Yinghuan Shi, and Yang Gao. Progressive cross-camera soft-label learning for semi-supervised person re-identification. *arXiv preprint arXiv:1908.05669*, 2019.
- [17] 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*, 2016.
- [18] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *NIPS*, 2016.
- [19] 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*, pages 480–496, 2018.
- [20] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.
- [21] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, 2018.
- [22] Jinlin Wu, Yang Yang, Hao Liu, Shengcai Liao, Zhen Lei, and Stan Z. Li. Unsupervised graph association for person re-identification. In *ICCV*, 2019.
- [23] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018.
- [24] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. In *CVPR*, 2017.
- [25] Qize Yang, Ancong Wu, and Wei-Shi Zheng. Deep semi-supervised person re-identification with external memory. In *ICME*, 2019.
- [26] Qize Yang, Hong-Xing Yu, Ancong Wu, and Wei-Shi Zheng. Patch-based discriminative feature learning for unsupervised person re-identification. In *CVPR*, pages 3633–3642, 2019.
- [27] Yao Zhai, Xun Guo, Yan Lu, and Houqiang Li. In defense of the classification loss for person re-identification. In *CVPRW*, 2019.
- [28] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Densely semantically aligned person re-identification. In *CVPR*, 2019.
- [29] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015.
- [30] 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.
- [31] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, 2017.
- [32] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *CVPR*, 2019.
- [33] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Learning to adapt invariance in memory for person re-identification. *arXiv preprint arXiv:1908.00485*, 2019.
- [34] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *ICCV*, 2019.
- [35] Xiangping Zhu, Xiatian Zhu, Minxian Li, Pietro Morello, Vittorio Murino, and Shaogang Gong. Intra-camera supervised person re-identification. *arXiv preprint arXiv:2002.05046*, 2020.

- [36] Xiangping Zhu, Xiatian Zhu, Minxian Li, Vittorio Murino, and Shaogang Gong. Intra-camera supervised person re-identification: A new benchmark. In *ICCVW*, 2019.