Intra-Camera Supervised Person Re-Identification

Xiangping Zhu[1,2] · Xiatian Zhu[3] · Minxian Li[4] · Pietro Morerio[2] · Vittorio Murino[2,5,6] · Shaogang Gong[4]

Abstract

Existing person re-identification (re-id) methods mostly exploit a large set of cross-camera identity labelled training data. This requires a tedious data collection and annotation process, leading to poor scalability in practical re-id applications. On the other hand unsupervised re-id methods do not need identity label information, but they usually suffer from much inferior and insufficient model performance. To overcome these fundamental limitations, we propose a novel person re-identification paradigm based on an idea of *independent* per-camera identity annotation. This eliminates the most time-consuming and tedious inter-camera identity labelling process, significantly reducing the amount of human annotation efforts. Consequently, it gives rise to a more scalable and more feasible setting, which we call *Intra-Camera Supervised (ICS)* person re-id, for which we formulate a Multi-tAsk mulTi-labEl (MATE) deep learning method. Specifically, MATE is designed for self-discovering the cross-camera identity correspondence in a per-camera multi-task inference framework. Extensive experiments demonstrate the cost-effectiveness superiority of our method over the alternative approaches on three large person re-id datasets. For example, MATE yields 88.7% rank-1 score on Market-1501 in the proposed ICS person re-id setting, significantly outperforming unsupervised learning models and closely approaching conventional fully supervised learning competitors.

Keywords Person re-identification · Intra-camera labelling · Cross-camera labelling · Multi-task learning · Multi-label learning

1 Introduction

Person re-identification (re-id) aims to retrieve the target identity class in detected person bounding box images captured by non-overlapping camera views (Gong et al. 2014; Prosser et al. 2010; Farenzena et al. 2010; Li et al. 2014; Zheng et al. 2013). It is a challenging task due to the non-rigid structure of human body, highly unconstrained appearance variation across cameras, and the low resolution and low

quality of the observations (Fig. 1a). While deep learning methods (Chen et al. 2017; Li et al. 2018b; Sun et al. 2018; Hou et al. 2019; Zheng et al. 2019; Zhou et al. 2019) have demonstrated remarkable performance advances, they rely on supervised model learning from a large set of cross-camera identity labelled training samples. This paradigm needs an exhaustive and expensive training data annotation process (Fig. 1b), dramatically lowering the usability while affecting the scalability of these methods for large scale deployment in real-world applications.

✉ Xiatian Zhu
  eddy.zhuxt@gmail.com

  Xiangping Zhu
  xiangping.zhu2010@gmail.com

  Minxian Li
  m.li@qmul.ac.uk

  Pietro Morerio
  pietro.morerio@iit.it

  Vittorio Murino
  vittorio.murino@univr.it

  Shaogang Gong
  s.gong@qmul.ac.uk

[1] Computer Vision Institute, School of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China

[2] Pattern Analysis and Computer Vision (PAVIS), Istituto Italiano di Tecnologia, Genova, Italy

[3] Vision Semantics Limited, London, UK

[4] Queen Mary University of London, London, UK

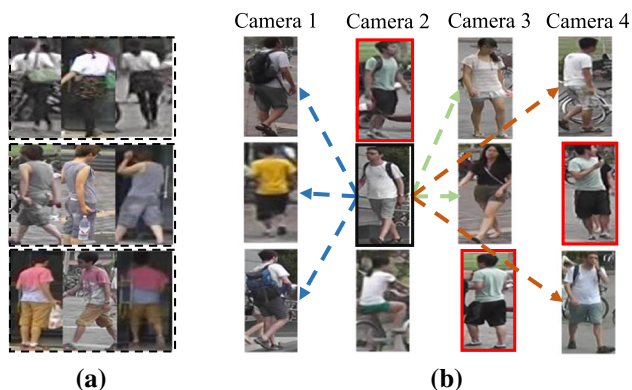[5] Ireland Research Center, Huawei Technologies Co. Ltd., Dublin, Ireland

[6] Department of Computer Science, University of Verona, Verona, Italy

Specifically, for constructing a conventional person re-id training dataset, human annotators usually need to annotate person identity labels both within individual camera views and across different camera views, and match a given person identity from one camera view with all the persons from other camera views (inter-camera person identity association). In particular, associating identity classes across camera views has a quadratic complexity with the number of both camera views and person identities (Fig. 1a). This would significantly increase the cost of creating conventional training dataset.

To quantify the annotation complexity, we consider that (1) there are N persons and M camera views, and (2) the cost of labelling every person is similar (average cost). To label one person in an *intra-camera annotation*, it requires to compare this person with all the other unlabelled persons and the labelling complexity is $O(N)$. So the labelling complexity for annotating all persons in one camera view is $O(N^2)$, and $O(MN^2)$ for all the camera views. For an *inter-camera annotation* (association), we start with the intra-camera labelling results. Given a person identity from one camera-view, an annotator needs to compare it exhaustively against the N identities from anyone of the other $M-1$ camera-views, i.e., $N(M-1)$ identities. This gives rise to a complexity of $O(N(M-1))$. To label N different persons, the annotation complexity is $O(N^2(M-1))$. As not all persons would appear in every camera view in most cases, this cross-camera view association needs to repeat for all M camera views, and the actual cost can vary according to the proportion of people reappearing in pairs of camera views.

Therefore the inter-camera annotation complexity is between two extremes: $O(N^2M)$ for exhaustive reappearing, and $O(M^2N^2)$ for zero reappearing.
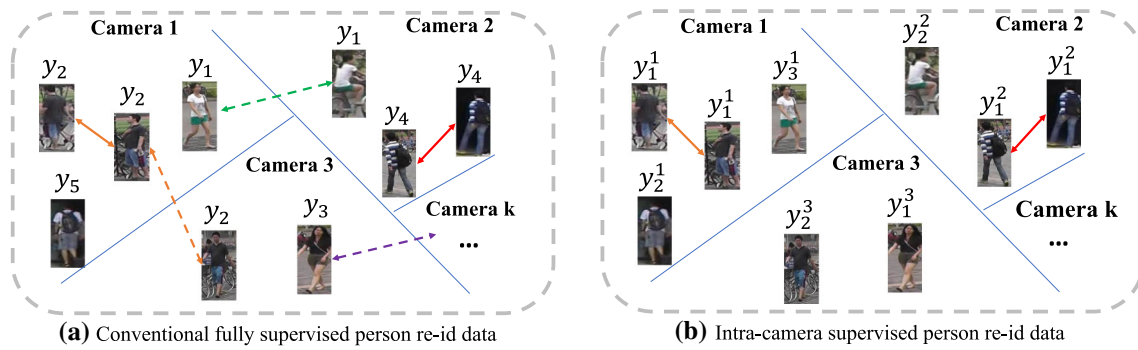
The problem of expensive training data collection has received significant attention. Representative attempts for minimising the annotation cost include:

(1) Domain generic feature design (Gray and Tao 2008; Farenzena et al. 2010; Zheng et al. 2015; Liao et al. 2015; Matsukawa et al. 2016),

(2) Unsupervised domain adaptation (Peng et al. 2016; Deng et al. 2018a; Wang et al. 2018; Lin et al. 2018; Zhong et al. 2018; Yu et al. 2019a; Chen et al. 2019),

(3) Unsupervised image/tracklet model learning (Wang et al. 2016a; Chen et al. 2018a; Lin et al. 2019; Li et al. 2019; Wu et al. 2020), and

(4) Weakly supervised learning (Meng et al. 2019). By hand-crafting generic appearance features with prior knowledge, the *first* paradigm of methods can perform re-id matching universally. However, their performances are often inferior due to limited knowledge encoded in such image representations. This can be addressed by transferring the labelled training data of a source dataset (domain), as demonstrated in the *second* paradigm of methods. Implicitly, these methods assume that the source and target domains share reasonably similar camera viewing conditions for ensuring sufficient transferable knowledge. The heavy reliance on the relevance and quality of source datasets (Zhu et al. 2019a) renders this approach less practically useful, since this assumption is often invalid. The *third* paradigm of methods is more scalable, as they need only unlabelled target domain data. While having high potential, unsupervised re-id methods usually yield the weakest performance, making them fail to meet the deployment requirements. In contrast, the *fourth* paradigm of methods considers a weakly supervised learning setting, where the person identity labels are annotated at the video level without fine-grained bounding boxes. Apart from insufficient re-id accuracy, this paradigm is mostly sensible only when such weak labels can be cheaply obtained from certain domain knowledge, which however is not generically accessible.

In this work, we suggest another novel person re-identification paradigm for scaling-up the model training process, called *Intra-Camera Supervised* (ICS) person re-id (Fig. 2b). As the name indicates, ICS eliminates the sub-process of cross-camera identity association during annotation, which is the majority component of the standard annotation cost. Under the ICS paradigm the training data involves only the intra-camera annotated identity labels with each camera view labelled *independently*. Importantly, as aforementioned, ICS naturally enables a parallel annotation process by camera views without labelling conflict due to no cross-camera identity association (Fig. 3b). This desirable merit is lacking in the conventional training data labelling due to the difficulty of obtaining disjoint labelling tasks, e.g. subsets of person identity classes without overlap (Fig. 3a). While being similar to the concurrent work (Meng et al.

Camera 1   Camera 2   Camera 3   Camera 4

(a)   **(b)**

Fig. 1 **a** Person re-identification challenges. Each triplet bounded by a dashed box shows the images of a single person from different camera views. **b** Illustration of manually associating identities across camera-views. The dashed arrow denotes the comparison between two identities. The associated identities are bounded with red boxes (Color figure online)

(a) Conventional fully supervised person re-id data

(b) Intra-camera supervised person re-id data

Fig. 2 Labels in person re-id data. **a** Conventional fully supervised training data needs both *per-camera* and *cross-camera* identity annotation in a unified class space. **b** Intra-camera supervised (ICS) training data only needs *per-camera* identity annotated *independently* in each

camera view with a separate class space. Camera-view index is encoded as superscript of identity label in ICS person re-id data. Solid and dashed arrows denote intra-camera and inter-camera association, respectively

2019) since they both consider explicitly the training data labelling process, our ICS paradigm however does *not* assume specific domain knowledge therefore it is more generally applicable.

To solve the ICS re-id problem, we propose a **Multi-tAsk mulTi-labEl** (MATE) deep learning model. Unlike the conventional fully supervised re-id methods using inter-camera identity labels, MATE is designed specially for overcoming two ICS challenges: (1) how to learn effectively from per-camera independently labelled training data, and (2) how to discover reliably the missing identity association across camera views. Specifically, MATE integrates two complementary learning components into a unified model: (a) *Per-camera multi-task learning* that separately learn individual camera views for modelling their specificity and the implicit shared information in a multi-task learning manner (Sect. 4.1). This assigns a specific network branch (i.e. a learning task) for modelling each camera view while constraining all the per-camera tasks to share a feature representation space. (b) *Cross-camera multi-label learning* that associates the identity labels across camera views in a multi-label learning strategy (Sect. 4.2). This is based on an idea of curriculum cyclic association that can associate reliably multiple cross-camera identity classes from self-discovered identity matches for multi-label model optimisation.

The *contributions* of this work are:

(1) We present a novel person re-identification paradigm for scaling up the model training process, dubbed as *Intra-Camera Supervised* (ICS) person re-id. ICS is characterised by *no* need for exhaustive cross-camera identity matching during training data annotation, whilst allowing naturally parallel labelling by camera views

(a)   **(b)**
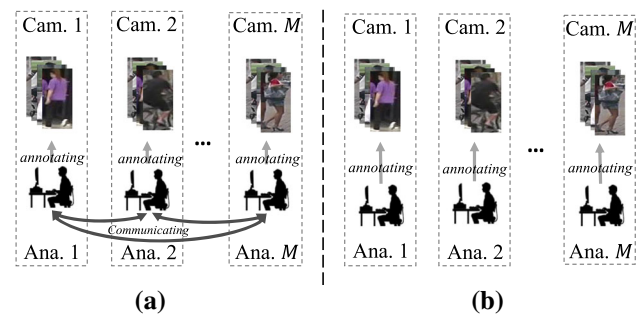
Fig. 3 Illustrations of data annotation process. **a** Conventional fully supervised person re-id vs. **b** ICS person re-id in the process of *training data collection*. Suppose each annotator needs to label the training data from a different camera view. In order to minimise the labelling conflict, an annotator may have to check if a person has been labelled or not by others. This gives rise to expensive communication costs, which is totally eliminated in the proposed ICS re-id paradigm, due to the independence nature between camera views

without conflict. Consequently, it makes the training data collection substantially cheaper and faster than the standard cross-camera identity labelling, therefore offering a more scalable mechanism to large re-id deployments.

(2) We formulate a *Multi-tAsk mulTi-labEl* (MATE) deep learning method for solving the proposed ICS person re-id problem. In particular, MATE combines the strengths of multi-task learning and multi-labelling learning in a unified framework to account for independent camera-specific identity label information and self-discovering their cross-camera association relationships concurrently. This represents a natural strategy for fully leveraging the ICS supervision with per-camera independent identity label spaces.

(3) Through extensive benchmarking and comparisons on the ICS variant of three large re-id datasets [Market-1501 (Zheng et al. 2015), DukeMTMC-reID (Zheng et al. 2017; Ristani et al. 2016), and MSMT17 (Wei

et al. 2018)], we demonstrate the cost-effectiveness advantages of the ICS re-id paradigm using our MATE model over the existing representative solutions including supervised learning, semi-supervised learning, unsupervised learning, unsupervised domain adaptation, and tracklet learning.

A *preliminary* version of this work was published in Zhu et al. (2019b). Compared with this earlier study, there are a number of key differences:

(i) This study presents a more comprehensive investigation into the proposed ICS person re-id paradigm in terms of training data annotation complexity, along with a comparison to the standard cross-camera identity labelling method. This provides a more accurate measurement of training data collection cost, revealing explicitly the intrinsic obstacles to scaling up model training as suffered by the conventional supervised learning re-id paradigm.

(ii) We propose a more principled *Multi-tAsk mulTi-labEl* learning method that can self-discover the cross-camera identity associations in a curriculum learning spirit. This improves dramatically the accuracy of cross-camera identity matching and therefore the final model generalisation, as compared to the earlier method. Besides, this new model performs unified end-to-end training without the need for two-stage learning as required in the earlier version.

(iii) We provide more comprehensive evaluations and analyses of the ICS person re-id for giving holistic and useful insights, in comparison to the existing alternative re-id paradigms.

2 Related Work

Supervised person re-id Most existing person re-id models are created by *supervised* learning methods on a separate set of cross-camera identity labelled training data (Wang et al. 2014b, 2016b; Zhao et al. 2017; Chen et al. 2017; Li et al. 2017; Chen et al. 2018b; Li et al. 2018b; Song et al. 2018; Chang et al. 2018; Sun et al. 2018; Shen et al. 2018a; Wei et al. 2018; Hou et al. 2019; Zheng et al. 2019; Zhang et al. 2019; Wu et al. 2019; Quan et al. 2019; Zhou et al. 2019). Relying on the strong supervision of cross-camera identity labelled training data, they have achieved remarkable performance boost. However, collecting such training data for each target domain is highly expensive, limiting their usability and scalability in real-world deployments at scales.

Semi-supervised person re-id A typical strategy for supervision minimisation is by semi-supervised learning. The key idea is to self-mine supervision information from unlabelled training data based on the knowledge learned from a small proportion of labelled training data. A few attempts have been made in this research direction (Figueira et al. 2013; Liu et al. 2014; Wang et al. 2016a; Xin et al. 2019). However, this paradigm not only suffers from significant performance degradation but also still needs a fairly large proportion of expensive cross-view pairwise labelling.

Weakly supervised person re-id Recently, Meng et al. (2019) propose a weakly supervised person re-id paradigm where the identity labels are annotated at the untrimmed video level. This setting makes more sense when such identity labels are readily available from certain domain knowledge which may be not generally provided. This is because, the major annotation cost of re-id training data comes from matching identity classes across camera views, rather than drawing person bounding boxes. Often, person images are directly detected from the raw videos by an on-the-shelf person detection model. Therefore, this paradigm is not sufficiently general.

Unsupervised person re-id Unsupervised model learning is an intuitive solution to avoid the need of exhaustively collecting a large number of labelled training data for every application domain. Early hand-crafted feature based unsupervised learning methods (Wang et al. 2014a; Kodirov et al. 2015, 2016; Khan and Bremond 2016; Ma et al. 2017; Ye et al. 2017; Liu et al. 2017) offer significantly inferior re-id matching performance, when compared to the supervised learning counterparts. Deep learning based methods (Lin et al. 2019; Wu et al. 2020) reduce this performance gap. Besides, there are two research lines on unsupervised re-id learning that become increasingly topical recently.

(1) *Unsupervised domain adaptation*
   The key idea of domain adaptation based methods (Wang et al. 2018; Fan et al. 2018; Peng et al. 2018; Yu et al. 2017; Zhu et al. 2017; Deng et al. 2018b; Zhong et al. 2018) is to explore the knowledge from the labelled data in *related* source domains with model adaptation on the unlabelled target domain data. Typical strategies include appearance style transfer (Zhu et al. 2017; Deng et al. 2018b; Chen et al. 2019), semantic attribute knowledge transfer (Peng et al. 2018; Wang et al. 2018), and progressive source appearance information adaptation (Fan et al. 2018; Yu et al. 2017). Although performing better than the earlier unsupervised learning methods, they require implicitly similar data distributions between the labelled source domain and the unlabelled target domain. This limits their scalability to arbitrarily diverse (unknown) target domains in real-world deployments.

(2) *Unsupervised tracklet learning* Instead of assuming transferable source domain training data, a small number of methods (Li et al. 2018a, 2019; Chen et al. 2018a; Wu et al. 2020) leverage the auto-generated tracklet data with rich spatio-temporal information for unsupervised re-id model learning. In many cases this is a feasible solution

as long as video data are available. However, it remains highly challenging to achieve good model performance due to noisy tracklets with unconstrained dynamics.

In this work, we introduce a new more scalable person re-id paradigm characterised by intra-camera supervised (ICS) learning, complementing the existing re-id scenarios as mentioned above. In comparison, ICS provides a superior trade-off between model accuracy and annotation cost, i.e. higher cost-effectiveness. This makes it a favourable choice for large scale re-id applications with high accuracy performance requirement and reasonably limited annotation budget.

3 Problem Formulation

We formulate the *Intra-Camera Supervised* (ICS) person re-identification problem. As illustrated in Fig. 2b, ICS only needs to annotate intra-camera person identity labels *independently*, whilst eliminating the most-expensive inter-camera identity association as required in the conventional fully supervised re-id setting.

Suppose there are M camera views in a surveillance camera network. For each camera view $p \in \{1, 2, \ldots, M\}$, we *independently* annotate a set of training images $\mathcal{D}^p = \{(\mathbf{x}_i^p, y_k^p)\}$ where each person image $\mathbf{x}_i^p$ is associated with an identity label $y_k^p \in \{y_1^p, y_2^p, \ldots, y_{N^p}^p\}$, and N^p is the total number of unique person identities in $\mathcal{D}^p$.[1] For clarity, we express the camera view index in the superscript due to the per-camera independent labelling nature in the ICS setting. By combining all the camera-specific labelled data $\mathcal{D}^p$, we obtain the entire training set as $\mathcal{D} = \{\mathcal{D}^1, \mathcal{D}^2, \ldots, \mathcal{D}^M\}$. For any two camera views p and q, their k-th person identities y_k^p and y_k^q usually describe two different people, i.e. they are two independent identity label spaces (Fig. 2b). This means exactly that the cross-camera identity association is *not* available, in contrast to the fully supervised re-id data annotation (Fig. 2a).

The ICS re-id problem presents a couple of new modelling challenges: (1) how to effectively exploit the per-camera person identity labels, and (2) how to automatically and reliably associate independent identity label spaces across camera views. The existing fully supervised re-id methods do not apply due to the need for identity annotation in a *single* label space across camera views. A new learning method tailored for the ICS setting is required to be developed.

4 Method

We introduce a novel ICS deep learning method, capable of conducting *M*ulti-t*A*sk mul*T*i-lab*E*l (*MATE*) model learning to fully exploit the independent per-camera person identity label spaces. In particular, MATE solves the aforementioned two challenges by integrating two complementary learning components into a unified solution: (i) *Per-camera multi-task learning* that assigns a separate learning task to each individual camera view for dedicatedly modelling the respective identity space (Sect. 4.1), (ii) *Cross-camera multi-label learning* that associates the independent identity label spaces across camera views in a multi-label strategy (Sect. 4.2). Combining the two capabilities with a unified objective function, MATE explicitly optimises their mutual compatibility and complementary benefits via end-to-end training. An overview of MATE is depicted in Fig. 4.

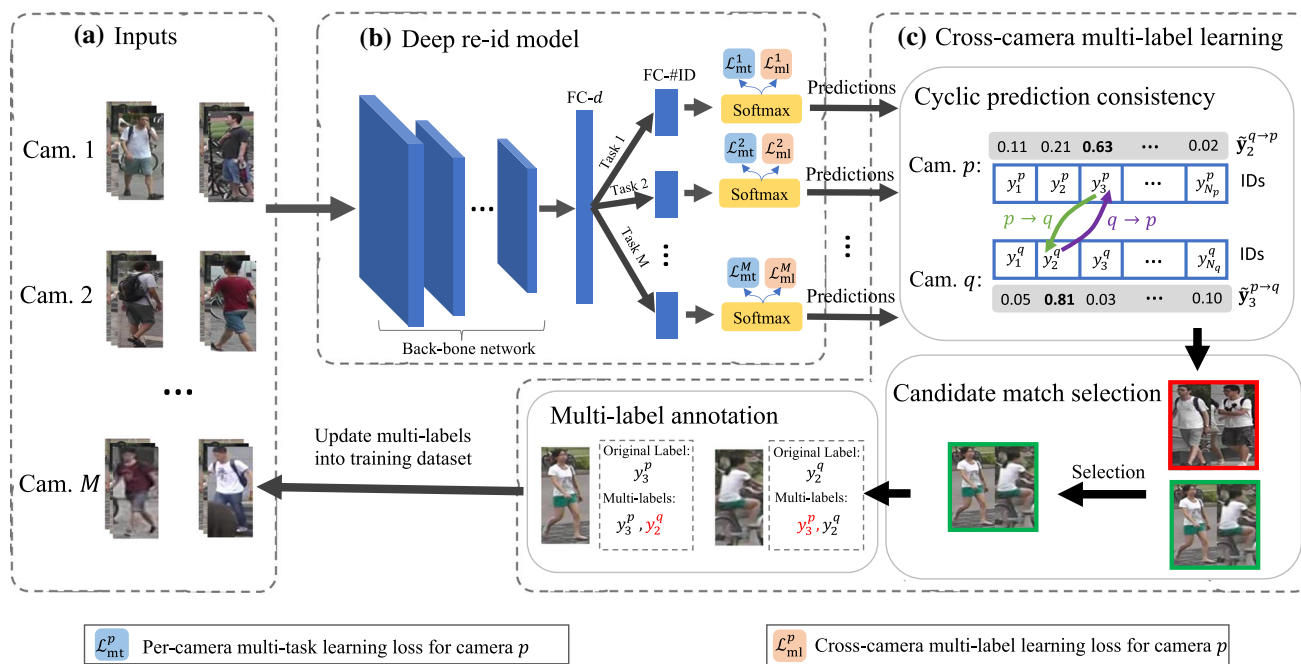### 4.1 Per-Camera Multi-Task Learning

To maximise the use of *multiple* camera-specific identity label spaces with some underlying correlation (e.g. partial identity overlap) in the ICS setting, multi-task learning is a natural choice for model design (Argyriou et al. 2007). This allows to not only mine the common knowledge among all the camera views, but also to improve per-camera model learning concurrently given augmented (aggregated) training data.

Specifically, given the nature of independent label spaces we consider each camera view as a separated learning task, all of which share a feature representation network for extracting the common knowledge in a multi-branch architecture design. One branch is in charge of a specific camera view. This forms *per-camera multi-task learning* in the ICS context. By such multi-task learning, our method can favourably derive a person re-id representation with *implicit* cross-camera identity discriminative capability, facilitating cross-camera identity association (Li et al. 2019). This is because during training, all the branches *concurrently* propagate the respective camera-specific identity label information through the shared representation network f_θ (Fig. 4b), leading to a *camera-generic* representation. This process is done by minimising the softmax cross-entropy loss.

Formally, for a training image $(\mathbf{x}_i^p, y_k^p) \in \mathcal{D}^p$ from camera view p, the softmax cross-entropy loss is used for formulating the training loss:

$$\mathcal{L}_{\mathrm{mt}}^p(i) = -\mathbb{1}(y_k^p)\log\Big(g^p\big(f_\theta(\mathbf{x}_i^p)\big)\Big) \tag{1}$$

where given the *camera-shared* feature vector $f_\theta(\mathbf{x}_i^p) \in \mathbb{R}^{d \times 1}$, the classifier $g^p(\cdot)$ for the camera view p predicts an identity class distribution in its own label space with

---

[1] We use i, j to denote image indexes, k, l, t to denote identity indexes and p, q to denote camera indexes.

Fig. 4 Overview of the proposed Multi-tAsk mulTi-labEl (*MATE*) deep learning method. **a** Given per-camera independently labelled training images, MATE aims to learn an identity discriminative feature representation model. This is achieved by designing two learning components: **b** *Per-camera multi-task* learning where we consider each individual camera view as a separate learning task with its own identity class space and optimise these camera-specific tasks on a common feature repre-

sentation (Sect. 4.1), and **c** *Cross-camera multi-task* learning where we self-discover the underlying identity matching relationships across camera views via curriculum cyclic association and design a multi-label optimisation algorithm to exploit these discovered cross-camera association information during model training. The two components are integrated together in a single MATE formulation, resulting in an end-to-end trainable model

N_p classes: $\mathbb{R}^{d \times 1} \rightarrow \mathbb{R}^{N_p \times 1}$. The Dirac delta function $\mathbb{1}(\cdot) : \mathbb{R} \rightarrow \mathbb{R}^{1 \times N_p}$ returns a one-hot vector with "1" at the specified index.

By aggregating the loss of training samples from all the camera views, we formulate the *per-camera multi-task learning* objective function as:

$$\mathcal{L}_{\mathrm{mt}} = \frac{1}{M} \sum_{p=1}^{M} \left(\frac{1}{B^p} \sum_{i=1}^{B^p} \mathcal{L}_{\mathrm{mt}}^p(i) \right) \qquad (2)$$

where B^p denotes the number of training images from the camera view p in a mini-batch.

4.2 Cross-Camera Multi-Label Learning

Cross-camera person appearance variation is a key challenge for re-id. Whilst this is implicitly modelled by the proposed multi-task learning as detailed above, the per-camera multi-task learning is still insufficient to fully capture the underlying identity correspondence relationships across camera-specific label spaces.

However, it is non-trivial to associate identity classes across camera views. One major reason is that a different set of persons may appear in a specific camera view, leading

to *no one-to-one identity matching* between camera views. Conceptually, this gives rise to a very challenging open-set recognition problem where a rejection strategy is often additionally required (Scheirer et al. 2013, 2014). Compared to generic object recognition in natural images, open-set modelling in re-id is more difficult due to small training data, large intra-class variation, subtle inter-class difference, and ambiguous visual observations of surveillance person imagery. Besides, existing open-set methods often assume accurately and completely labelled training data, and the unseen classes only in model test. In contrast, we need to discover cross-camera identity correspondences during training with small (unknown) overlap across different spaces.

This is hence a harder learning scenario with a higher risk of error propagation from noisy cross-camera association. An intuitive solution for open-set recognition is to find an operating threshold, e.g. by Extreme Value Theory (De Haan and Ferreira 2007) based statistical analysis. This relies on optimal *supervised* model learning from a sufficiently large training dataset, which however is unavailable in the ICS setting.

To circumvent the above problems, we design a *cross-camera multi-label learning* strategy for robust cross-camera identity association. This is realised by (i) designing a *curriculum cyclic association constraint* to find reliable

cross-camera identity association, and (ii) forming a *multi-label learning algorithm* to incorporate the self-discovered cross-camera identity association into discriminative model learning (Fig. 4c).

4.2.1 Curriculum Cyclic Association

For more reliable identity association across camera views, we form a *cyclic prediction consistency* constraint. Specifically, given an identity class $y_k^p \in \{y_1^p, y_2^p, \ldots, y_{N^p}^p\}$ from a camera view $p \in \{1, 2, \ldots, M\}$, we need to find if a true matching identity (i.e. the same identity) exists in another camera view q. We achieve this in the following process.

(i) We first project all the images of each person identity y_k^p from camera view p to the classifier branch of camera view q to obtain a *cross-camera prediction* $\tilde{\mathbf{y}}_k^{p \to q}$ via averaging as:

$$\tilde{\mathbf{y}}_k^{p \to q} = \frac{1}{S_k^p} \sum_{i=1}^{S_k^p} g^q \left(f(\mathbf{x}_i^p) \right) \in \mathbb{R}^{N_q \times 1}, \tag{3}$$

where S_k^p is the number of images from identity y_k^p. Each element of $\tilde{\mathbf{y}}_k^{p \to q}$, denoted as $\tilde{\mathbf{y}}_k^{p \to q}(l)$, means the identity class matching probability at which y_k^p (an identity from camera view p) matches y_l^q (an identity from camera view q) in a cross-camera sense.

(ii) We then nominate the person identity y_{l*}^q from camera view q with the maximum likelihood probability as the candidate matching identity:

$$l^* = \arg\max_l \tilde{\mathbf{y}}_i^{p \to q}(l), \ l \in \{1, 2, \ldots, N_q\}. \tag{4}$$

With such one-way $(p \to q)$ association alone, the matching accuracy should be not satisfactory since it cannot handle the cases of *no-true-match* as typical in the ICS setting. To boost the matching robustness and correctness, we further design a *curriculum cyclic association* constraint.

(iii) Specifically, in an opposite direction of the above steps, we project all the images of identity y_{l*}^q from camera view q to the classifier branch of camera view p in a similar way as Eq. (3), and obtain the best candidate matching identity y_{t*}^p with Eq. (4). Given this back-and-forth matching between camera view p and q, we subsequently filter the above candidate pair (y_k^p, y_{l*}^q) by a cyclic constraint as:

$$(y_k^p, y_{l*}^q) \begin{cases} \text{is a candidate match,} & \text{if } y_{t*}^p = y_k^p, \\ \text{is not a candidate match,} & \text{otherwise.} \end{cases} \tag{5}$$

This removes non-cyclic association pairs. While being more reliable, it is observed that only the cyclic association in Eq. (5) is not sufficiently strong for *hard* cases (e.g. different people with very similar clothing appearance), leading to false association.

(iv) To overcome this problem, inspired by the findings of cognitive study which suggest a better learning strategy is to start *small* (Elman 1993; Krueger and Dayan 2009), we design a curriculum association constraint. It is based on the cross-camera identity matching probability. Formally, we define a cyclic association degree as:

$$\psi_{k \Leftrightarrow l^*}^{p \Leftrightarrow q} = \tilde{\mathbf{y}}_k^{p \to q}(l^*) \cdot \tilde{\mathbf{y}}_{l^*}^{q \to p}(k) \tag{6}$$

which measures the joint probability of a cyclic association between two identities y_k^p and $y_{l^*}^q$. Given this unary measurement, we can deploy a *curriculum threshold* $\tau \in [0, 1]$ for selecting candidate matching pairs via:

$$\text{Cyclic } (y_k^p, y_{l^*}^q) \begin{cases} \text{is a match,} & \text{if } \psi_{k \Leftrightarrow l^*}^{p \Leftrightarrow q} > \tau, \\ \text{is not a match,} & \text{otherwise.} \end{cases} \tag{7}$$

This filtering determines if a cyclically associated identity pair (y_i^p, y_{k*}^q) will be considered as a match.

Curriculum threshold The design of the curriculum threshold τ has a crucial influence on the quality of cross-camera identity association. In the spirit of curriculum learning, we consider τ as an annealing function of the model training time to enable a progressive selection. Meanwhile, we need to take into account that the magnitude of maximum prediction usually increases along the training process as the model gets more mature. Taking these into consideration, we formulate the curriculum threshold as:

$$\tau^r = \min \left(\tau^u, \ \tau^l + \frac{r}{R - 1}(1 - \tau^l) \right) \tag{8}$$

where r specifies the current training round, with a total of R rounds. We maintain two thresholds: τ^u, which denotes the upper bound, and τ^l, which denotes the lower bound. Both of these two thresholds can be estimated by cross-validation.

Summary We perform the above curriculum cyclic association process for every camera view pairs, which outputs a set of associated identity pairs across camera views. This self-discovered pairwise information will be used to improve model training as detailed in the following.

4.2.2 Multi-Label Learning

To leverage the above identity association results for improving model discriminative learning, we introduce a multi-label learning scheme in a cross-camera perspective. It consists of (i) multi-label annotation and (ii) multi-label training.

(i) *Multi-label annotation.* For easing presentation and understanding, we assume two camera views, and it is straightforward to extend to more camera views. Given an associated identity pair (y_k^p, y_{l*}^q) obtained as above, we annotate all the images X_i^p of y_i^p from camera view p with an extra label y_{l*}^q of camera view q. We do the same for all the images X_{l*}^q of y_{l*}^q in an inverse direction. Both image sets are therefore annotated with the same two identity labels, i.e. these images are associated. See an illustration example in Fig. 4c. Given M camera views, for each identity y_k^p we perform at most $M-1$ times such annotation whenever a cross-camera association is found, resulting in a multi-label set $Y_i^p = \{y_k^p, y_{l*}^q, \ldots\}$ for X_i^p, with the cardinality $1 \leq |Y_i^p| \leq M$. When $|Y_i^p| = 1$, it means no cross-camera association is obtained. When $|Y_i^p| = M$, it means an identity association is found in every other camera view.

(ii) *Multi-label training.* Given such cross-camera multi-label annotation, we then formulate a multi-label training objective for an image $\mathbf{x}_i^p$ as

$$\mathcal{L}_{ml}^p(i) = \frac{1}{|Y_i^p|} \sum_{y^c \in Y_i^p} -\mathbb{1}(y^c)\log\Big(g^c\big(f_\theta(\mathbf{x}_i^p)\big)\Big) \qquad (9)$$

where c indices the camera view of Y_i^p with the corresponding identity label simplified as y^c. For mini-batch training, we design the cross-camera multi-label learning objective as:

$$\mathcal{L}_{ml} = \frac{1}{B} \sum_{i,p} \mathcal{L}_{ml}^p(i) \qquad (10)$$

which averages the multi-label training loss of all the B number of training images in a mini-batch.

Remarks It is noteworthy to point out that, in contrast to the conventional single-task multi-label learning (Tsoumakas and Katakis 2007), we jointly form multi-label learning and multi-task learning in a unified framework, with a unique objective of associating different label spaces and merging the independently annotated labels with the same semantics.

4.3 Final Objective Loss Function

By combining per-camera multi-task (Eq. (2)) and cross-camera multi-label (Eq. (10)) learning objectives, we obtain the final model loss function as:

$$\mathcal{L} = \mathcal{L}_{mt} + \lambda \mathcal{L}_{ml}, \qquad (11)$$

where the weight parameter $\lambda \in [0, 1]$ is to trade-off the two loss terms. With this formula as model training supervision, our method can effectively learn discriminative re-id model using both camera-specific identity label spaces available under the ICS setting ($\mathcal{L}_{mt}$) and cross-camera identity association self-discovered by MATE itself ($\mathcal{L}_{ml}$) concurrently. The MATE model training process is summarised in Algorithm 1.

Algorithm 1 The MATE model training procedure.

Input: Intra-camera independently labelled training data;
Output: A trained person re-id model;
Model training:
  for $r = 1$ **to** R **do**:
    Calculate the curriculum threshold τ^r;
    Cross-camera identity association as in Eqs. (3)–(7);
    for $e = 1$ **to** *epoch_number* **do**:
      for $t = 1$ **to** *per-epoch mini-batch number* **do**:
        Feed forward a mini-batch of training images;
        Compute learning loss using Eq. (11);
        Update the network model by back-propagation;
      end for
    end for
  end for

---

5 Experiments

Datasets Due to *no* existing re-id datasets for the proposed scenario, we introduced three ICS re-id benchmarks. We simulated the ICS identity annotation process on three existing large person re-id datasets, Market-1501 (Zheng et al. 2015), DukeMTMC-reID (Ristani et al. 2016; Zheng et al. 2017) and MSMT17 (Wei et al. 2018). Specifically, for the training data of each dataset, we *independently* perturbed the original identity labels for every individual camera view, and ensured that the same class labels of any pair of different camera views correspond to two unique persons (i.e. no labelled cross-camera association). We used the same original test data of each dataset for model performance evaluation.
Performance metrics Following the common person re-id works, the Cumulative Matching Characteristic (CMC) and mean Average Precision (mAP) metrics were used for model performance measurement.

Table 1 Benchmarking the ICS person re-id performance

Dataset	Market-1501			
Metric (%)	R1	R10	R20	mAP
MCST	34.9	60.1	69.3	16.7
EPCS	42.6	64.6	71.2	19.6
PCMT	78.4	93.1	95.7	52.1
MATE (Ours)	88.7	97.1	98.2	71.1
Dataset	DukeMTMC-reID			
Metric (%)	R1	R10	R20	mAP
MCST	25.0	50.1	58.8	16.3
EPCS	38.8	58.9	64.6	22.1
PCMT	65.2	81.1	85.6	44.7
MATE (Ours)	76.9	89.6	92.3	56.6
Dataset	MSMT17			
Metric (%)	R1	R10	R20	mAP
MCST	12.1	26.3	33.0	4.8
EPCS	16.8	31.5	37.4	5.4
PCMT	39.6	59.6	65.7	15.9
MATE (Ours)	46.0	65.3	71.1	19.1

Implementation details The ImageNet pre-trained ResNet-50 (He et al. 2016) was selected as the backbone network of our MATE model. As shown in Fig. 4, each branch in MATE was formed by a fully connected (FC) classification layer. We set the dimension of the re-id feature representation to 512. The person images were resized to 256×128 in pixel. The standard stochastic gradient descent (SGD) optimiser was adopted. The initial learning rate of the backbone network and classifiers were set to 0.005 and 0.05, respectively. We set a total of 10 rounds to anneal the curriculum threshold τ (Eq. (7)), with each round covering 20 epochs (except the last round where we trained 50 epochs to guarantee the convergence). We empirically estimated $\tau^l = 0.5$ (the lower bound of τ) and $\tau^u = 0.95$ (the upper bound of τ) for Eq. (8). In order to balance the model training across camera views, we randomly selected from each camera the same number of images, i.e. 4 images, per identity and the same number of identities, i.e. 2 identities, to construct a mini-batch. Unless stated otherwise, we set the loss weight $\lambda = 0.5$ for Eq. (11). In test, the Euclidean distance was applied to the camera-generic feature representations for re-id matching.

5.1 Benchmarking the ICS Person Re-ID

Since there is no dedicated methods for solving the proposed ICS person re-id problem, we formulated and benchmarked three baseline methods based on the generic learning algorithms:

1. *Multi-Camera Single-Task* (MCST) learning (Fig. 5a): given no identity association across camera views, we simply consider that identity classes from different camera views are distinct people and merge all the per-camera label spaces into a joint space cumulatively. This enables the conventional supervised model learning based on identity classification. We therefore train a single re-id model, as in the common supervised learning paradigm. At test time, we extract the re-id feature vectors and apply the Euclidean distance as the metrics for re-id matching.
2. *Ensemble of Per-Camera Supervised* (EPCS) learning (Fig. 5b): without inter-camera identity labels, for each camera view we train a separate re-id model with its own single-camera training data. During deployment, given a test image we extract the feature vectors of all the per-camera models, concatenate them into a single representation vector, and utilise the Euclidean distance as the matching metrics for re-id.
3. *Per-Camera Multi-Task* (PCMT) learning (Fig. 5c): while being a variant of our MATE model *without* the cross-camera multi-label learning component, we simultaneously consider it as a baseline due to the use of the multi-task learning strategy.

To implement fairly the baseline learning methods, we used the same backbone ResNet50 as our method, a widely used architecture in the re-id literature. We trained each of these models with the softmax cross-entropy loss function in their respective designs.

Results We compared our MATE model with the three baseline methods in Table 1. Several observations can be pointed:

1. Concatenating simply the per-camera identity label spaces, MCST yields the weakest re-id performance. This is not surprised because there is a large (unknown) proportion of duplicated identities but mistakenly labelled with different classes, misleading the model training process.
2. The above problem can be addressed by independently exploiting camera-specific identity class annotations, as EPCS does. This method does produce better re-id model generalisation consistently. However, the over accuracy is still rather low, due to the incapability of leveraging the shared knowledge between camera views and mining the inter-camera identity matching information.
3. To address this cross-camera association issue, PCMT provides an implicit solution and significantly improves the model performance.
4. Moreover, the proposed MATE model further boosts the re-id matching accuracy by explicitly associating the identity classes across camera views in a reliable formulation. This verifies the efficacy of our model in capitalising such cheaper and more scalable per-camera identity labelling.

(a) MCST   **(b)** EPCSL   **(c)** PCMT
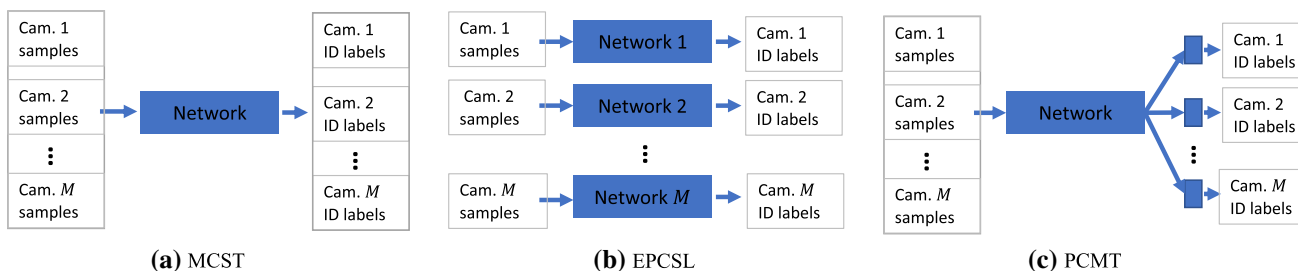
Fig. 5 Three baseline learning methods for ICS person re-id: **a** Multi-Camera Single-Task (MCST) learning. **b** Ensemble of Per-Camera Supervised (EPCS) Learning. **c** Per-Camera Multi-Task (PCMT) learning

(a) MCST   **(b)** EPCSL   **(c)** PCMT   **(d)** MATE
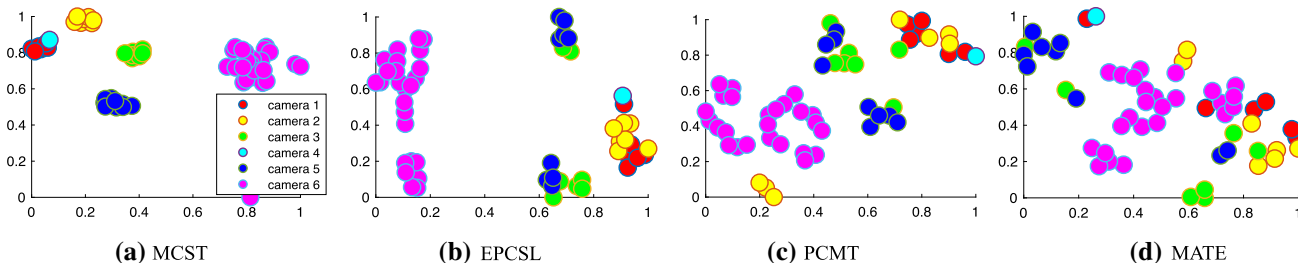
Fig. 6 Feature distribution visualisation of a randomly selected person identity appearing under all the six camera views of the Market-1501 dataset. This is made by t-SNE (Maaten and Hinton 2008). Camera views are colour-coded. Best viewed in colour (Color figure online)

Table 2 Comparative evaluation of representative person re-id paradigms in the model training *supervision* perspective

Supervision	Method	Market-1501				DukeMTMC-reID				MSMT17			
		R1	R10	R20	mAP	R1	R10	R20	mAP	R1	R10	R20	mAP
None	RKSL[†]	34.0	–	–	11.0	–	–	–	–	15.4	–	–	4.3
	ISR[†]	40.3	–	–	14.3	–	–	–	–	21.5	–	–	6.1
	DIC[†]	50.2	–	–	22.7	–	–	–	–	22.8	–	–	7.0
	BUC	66.2	84.5	–	38.3	47.4	68.4	–	27.5	–	–	–	–
	TSSL	71.2	–	–	43.3	62.2	–	–	38.5	–	–	–	–
Tracking	TAUDL	63.7	–	–	41.2	61.7	–	–	43.5	–	–	–	–
	UTAL	69.2	85.5	89.7	46.2	62.3	80.7	84.4	44.6	31.4	51.0	58.1	13.1
Source domain	CAMEL	54.5	–	–	26.3	–	–	–	–	–	–	–	–
	TJ-AIDL	58.2	–	–	26.5	44.3	–	–	23.0	–	–	–	–
	CR-GAN	59.6	–	–	29.6	52.2	–	–	30.0	–	–	–	–
	MAR	67.7	–	–	40.0	67.1	–	–	48.0	–	–	–	–
	ECN	75.1	91.6	–	43.0	63.3	80.4	–	40.4	30.2	46.8	–	10.2
Intra-camera	MATE (Ours)	88.7	97.1	98.2	71.1	76.9	89.6	92.3	56.6	46.0	65.3	71.1	19.1
Cross-camera (semi)	ResNet50*	66.1	–	–	42.1	50.0	–	–	30.3	–	–	–	–
	WRN50*	65.8	–	–	42.2	49.4	–	–	30.9	–	–	–	–
	MVC	72.2	–	–	49.6	52.9	–	–	33.6	–	–	–	–
Cross-camera	HA-CNN	91.2	–	–	75.7	80.5	–	–	63.8	–	–	–	–
	SGGNN	92.3	–	–	82.8	81.1	–	–	68.2	–	–	–	–
	PCB	93.8	–	–	81.6	83.3	–	–	69.2	68.2	–	–	40.4
	JDGL	94.8	–	–	86.0	86.6	–	–	74.8	77.2	–	–	52.3
	OSNet	94.8	–	–	84.9	88.6	–	–	73.5	78.7	–	–	52.9

[†]Results from Yu et al. (2019b)
*Results from Xin et al. (2019)

To further examine the model performance, in Fig. 6 we visualised the feature distributions of a randomly selected person identity with images captured by all the camera views of Market-1501. It is shown that the feature points of our model present the best camera-invariance property, qualitatively validating the superior re-id performance over other competitors.

5.2 Comparing Different Person Re-ID Paradigms

As a novel re-id person scenario, it is informative and necessary to compare with other existing scenarios in the problem-solving and supervision cost perspectives. To this end, we compared ICS with existing representative re-id paradigms in an increasing order of training supervision cost:

1. *Unsupervised learning* (no supervision): RKSL (Wang et al. 2016a), ISR (Lisanti et al. 2014), DIC (Kodirov et al. 2015), BUC (Lin et al. 2019), and TSSL (Wu et al. 2020);
2. *Tracking data modelling*: TAUDL (Li et al. 2018a) and UTAL (Li et al. 2019);
3. *Unsupervised domain adaptation* (source domain supervision): CAMEL (Yu et al. 2017), TJ-AIDL (Wang et al. 2018), CR-GAN (Chen et al. 2019), MAR (Yu et al. 2019b), and ECN (Zhong et al. 2019);
4. *Semi-supervised learning* (cross-camera supervision at small size): ResNet50 (He et al. 2016), WRN50 (Zagoruyko and Komodakis 2016), and MVC (Xin et al. 2019);
5. *Conventional fully supervised learning* (cross-camera supervision): HA-CNN (Li et al. 2018b), SGGNN (Shen et al. 2018b), PCB (Sun et al. 2018), JDGL (Zheng et al. 2019), and OSNet (Zhou et al. 2019).

Table 2 presents a comprehensive comparative evaluation of different person re-id paradigms in terms of the model performance *vs.* supervision requirement. We highlight the following observations:

1. Early unsupervised learning re-id models (RKSL, ISR, DIC), which rely on hand-crafted visual feature representations, often yield very limited re-id matching accuracy.
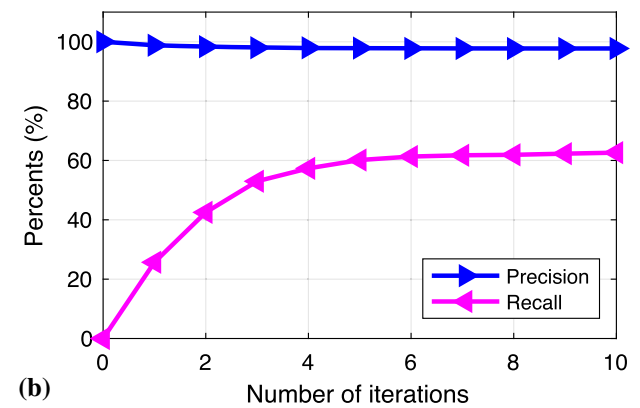
Table 3 Evaluating the model components of MATE: Per-Camera Multi-Task (PCMT) learning, Cross-Camera Multi-Label (CCML) learning, and Curriculum Thresholding (CT)

| Component	R1	R10	R20	mAP
PCMT	78.4	93.1	95.7	52.1
PCMT+CCML	85.3	96.2	97.6	65.2
PCMT+CCML+CT (full)	88.7	97.1	98.2	71.1

Dataset: Market-1501
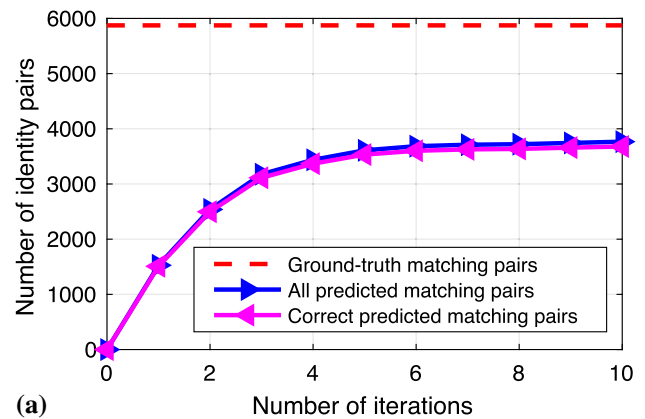
Fig. 7 Dynamic statistics of cross-camera identity association over the training rounds. Dataset: Market-1501. **a** The number of ground-truth matching pairs (*ground-truth pairs*), the number of all predicted matching pairs (*all predicted pairs*), and the number of correctly predicted matching pairs (*correct predicted pairs*). **b** The *precision* and *recall* of all predicted matching pairs

While deep learning clearly improves the performance as shown in BUC and TSSL, the results are still largely unsatisfactory.

2. By exploiting tracking information including spatio-temporal object appearance continuity, TAUDL and UTAL further improve the model generalisation.
3. Unsupervised domain adaptation is another classical approach to eliminating the tedious collection of labelled training data per domain. The key idea is knowledge transfer from a source dataset (domain) with cross-camera labelled training samples. This strategy continuously pushes up the matching accuracy. It has the clear limitation that a *relevant* labelled source domain is assumed which however is not always guaranteed in practice.
4. While semi-supervised learning enables label reduction, the model performance remains unsatisfactory and is relatively inferior to unsupervised domain adaptation. This paradigm relies on expensive cross-camera identity annotation despite at smaller sizes.

(a) Iteration 0 **(b)** Iteration 1 **(c)** Iteration 5 **(d)** Iteration 10 **(e)** Fully supervised training
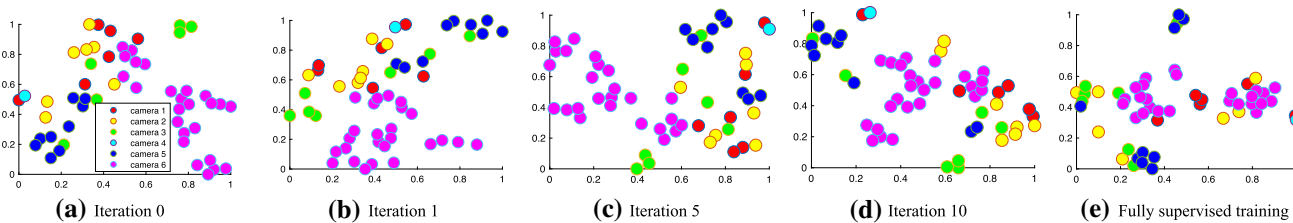
Fig. 8 **a–d** The feature distribution evolution of a set of multi-camera images from a single random person over the training rounds, in comparison to **e** the feature distribution by supervised learning. Iteration 0 indicates the initial feature space before training starts. Dataset: Market-1501 Best viewed in colour (Color figure online)

(a) Iteration 0 **(b)** Iteration 1 **(c)** Iteration 5 **(d)** Iteration 10 **(e)** Fully supervised training
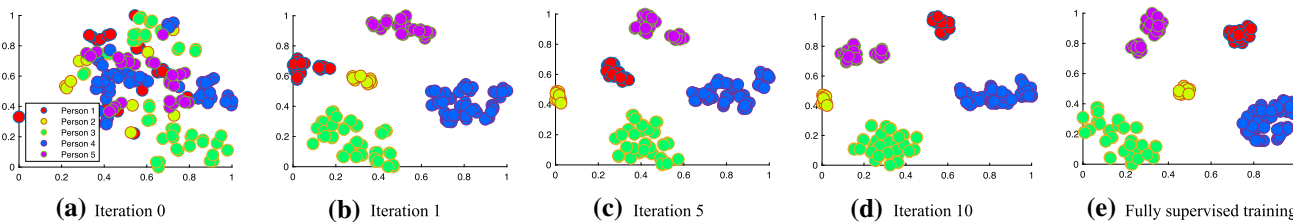
Fig. 9 **a–d** The feature distribution evolution of multi-camera images from five random persons over the training rounds, in comparison to **e** the feature distribution by supervised learning. Iteration 0 indicates the initial feature space before training starts. Dataset: Market-1501 Best viewed in colour (Color figure online)

5. With full cross-camera identity label supervision, supervised learning methods produce the best re-id performance among all the paradigms. However, the need for cross-camera identity association leads to very high labelling cost per domain, restricting significantly its scalability in realistic large scale applications typically with limited annotation budgets.

6. The ICS re-id is proposed exactly for solving this low cost-effectiveness limitation of the conventional supervised learning re-id paradigm, without the expensive cross-camera identity association labelling. Despite much weaker supervision, MATE can approach the performance of the latest supervised learning re-id methods on Market-1501. However, the performance gap on the largest dataset MSMT17 is still clearly bigger, suggesting a large room for further ICS re-id algorithm innovations.

5.3 Further Evaluation of Our Method

We conducted a sequence of in-depth component evaluations for the MATE model on the Market-1501 dataset.

5.3.1 Ablation Study

We started by evaluating the three components of our MATE model: *Per-Camera Multi-Task* (PCMT) learning, *Cross-Camera Multi-Label* (CCML) learning, and *Curriculum Thresholding* (CT). The results in Table 3 show that:


(a) **(b)** **(c)**
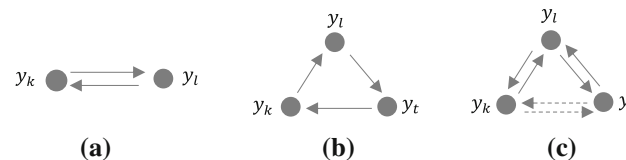
Fig. 10 Illustration of three methods for identity cyclic association across camera views: **a** Association across two camera views, adopted by MATE; **b** Association across three camera views; **c** Transitive association across three camera views. y_k, y_l, y_t are three identities from different camera views. Sold arrow denotes the correspondence relation discovered as in Sect. 4.2, and dashed arrow in (**c**) denotes transitive association. Both (**b**) and (**c**) can be further extended to more camera views

(1) Using the PCMT component alone, the model can already achieve fairly strong re-id matching performance, thanks to the ability of learning implicitly cross-camera feature representation via a specially designed multi-task inference structure.

(2) Adding the CCML component significantly boosts the accuracy, verifying the capability of our cross-camera identity matching strategy in discovering the underlying image pairs.

(3) With the help of CT, a further performance gain is realised, validating the idea of exploiting curriculum learning and the design of our curriculum threshold.

As a key performance contributor, we further examined CCML by evaluating its essential part—cross-camera identity association. To this end, we tracked the statistics of self-discovered identity pairs across camera views over the
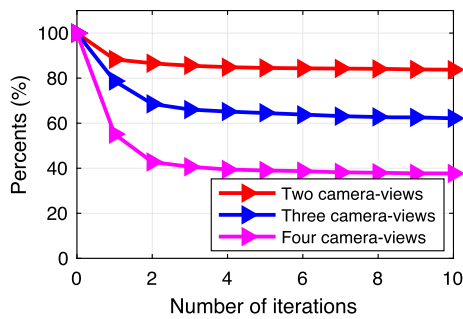
Fig. 11 Comparing the association precision when varying numbers of camera views are involved in cyclic consistent association

training rounds, including the precision and recall measurements. It is shown in Fig. 7 that our model can mine an increasing number of identity association pairs whilst maintaining very high precision which therefore well limits the risk of error propagation and its disaster consequence. This explains the efficacy of our cross-camera multi-label learning. On the other hand, while failing to identify around 40% identity pairs, our model can still achieve very competitive performance as compared to fully supervised learning models. This suggests that our method has already discovered the majority of re-id discrimination information from the associated identity pairs, missing only a small fraction embedded in those hard-to-match pairs. In this regard, we consider the proposed model is making a satisfactory trade-off between identity association error and knowledge mining. To check the impact of cross-camera identity association together with per-camera learning, we visualised the change of feature distribution during training. For a set of multi-camera images from a single person, it is observed in Fig. 8 that they are associated gradually in the re-id feature space, reaching a similar distribution as in the supervised learning case. For a set of images from five random persons,

our model enables them to be gradually pushed away, as shown in Fig. 9. These observations are in line with the numerical performance evaluation above.

Associative scope Conceptually, the proposed concept of cyclic consistent association can be extended to three or more camera views. An example for three camera views is illustrated in Fig. 10b. For a more focused evaluation, we analysed this aspect without curriculum threshold. We considered 2, 3, and 4 camera views involved during association. We obtained Rank-1/mAP rates of 85.3%/65.2%, 83.5%/64.2%, and 80.7%/58.9%, respectively. This result shows that the more camera views involved, the lower model performance obtained. The plausible reason is that the negative effect of error propagation would be amplified when additional camera views are added into the associating cycle. This is clearly reflected in the comparison of association precision, as shown in Fig. 11.

Transitive association As shown in Fig. 10c, transitive association means that if two identities (y_k and y_t) are both associated with another identity (y_l) in a cross-camera sense, then the two identities y_k and y_t should be also associated. In MATE, the transitive association is implicitly considered. More specifically, when y_k and y_t both are pulled close towards y_l concurrently, y_k and y_t will be made close in feature space during training, i.e. y_k and y_t are associated. This transitive association can be further extended to 4 or more camera views. To verify the above analysis, we evaluated the effect of explicitly exploiting the transitivity information in training MATE. We obtained 88.9%/71.2% in R1/mAP, similar to the performance of 88.7%/71.1% when it is implicitly utilised. In design, we finally choose to implicitly mine such transitive relations for reduced model complexity.

5.3.2 Hyper-Parameter Analysis

We examined the performance sensitivity of three parameters of MATE: the loss weight λ (default value 0.5) in Eq. (11), the lower (default value 0.5) and upper (default value 0.95) bound of curriculum threshold in Eq. (8). We evaluated each individual parameter by varying its value while setting all the others to their default values. Figure 12 shows that all these parameters have a wide range of satisfactory values in terms of performance. This suggests the ease and convenience of setting up model training and good accuracy stability of our method.
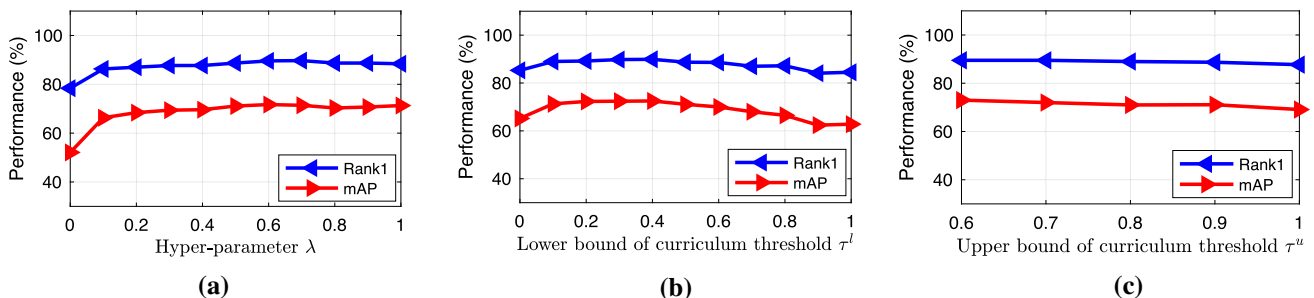
Fig. 12 Hyper-parameter analysis: **a** the loss weight λ in Eq. (11), the **b** lower and **c** upper bound of curriculum threshold in Eq. (8). Dataset: Market-1501

5.4 Intra-Camera Annotation Cost

We conducted a controlled data annotation experiment to annotate intra-camera person identity labels on the MSMT17 dataset (Wei et al. 2018). Specifically, we annotated the identity labels of person images in a camera-independent manner, with the original identity information discarded. Due to the nature of per-camera person labelling, the entire identity space is split into multiple independent, smaller spaces. This allows us to decompose the labelling task easily and enable multiple annotators to conduct the labelling job in parallel without any interference and conflict among them. These merits reduce significantly the annotation cost.

We provide a quantitative comparison on the annotation costs between ICS and a conventional fully supervised person re-id setting. This experiment was performed on a subset of MSMT17. Specifically, we randomly selected up to 50 persons from each camera-view which gives rise to a total of 714 identities. We asked three annotators to label the images using the same labelling tool we developed. The labelling costs of ICS and the fully supervised setting are respectively 2.5 and 8 person-days. This empirical validation is largely consistent with our annotation cost complexity analysis provided in the Introduction section. This experiment demonstrates that our ICS setting is significantly more efficient and more scalable by reducing the annotation complexity and costs.

In terms of performance, our method achieves a Rank-1/mAP rate of 46.0%/19.1%, *vs.* 78.7%/52.9% by the best supervised learning model OSNet, whilst clearly outperforming all unsupervised, tracking, domain adaptation based alternatives (cf. Table 1). This is an encouraging preliminary effort of intra-camera supervised person re-id, with a good improvement space remaining in algorithm innovation.

6 Conclusions

In this work, we presented a novel person re-identification paradigm, i.e. intra-camera supervised (ICS) learning, characterised by training re-id models with only per-camera independent person identity labels, but no the conventional cross-camera identity labelling. The key motivation lies in eliminating the tedious and expensive process of manually associating identity classes across every pair of camera views in a surveillance network, which makes the training data collection too costly to be affordable in large real-world application. To address the ICS re-id problem, we formulated a Multi-tAsk mulTi-labEl (MATE) learning model capable of fully exploiting per-camera re-id supervision whilst simultaneously self-discovering cross-camera identity association. Extensive evaluations were conducted on three re-id benchmarks to validate the advantages of the proposed MATE model over the state-of-the-art alternative methods in the proposed ICS learning setting. Detailed ablation analysis is also provided for giving insights on our model design. We conducted extensive comparative evaluations to demonstrate the cost-effectiveness advantages of the ICS re-id paradigm over a wide range of existing representative re-id settings and the performance superiority of our MATE model over alternative learning methods.

References

Argyriou, A., Evgeniou, T., & Pontil, M. (2007). Multi-task feature learning. In *Advances in neural information processing systems* (pp 41–48).

Chang, X., Hospedales, T. M., & Xiang, T. (2018). Multi-level factorisation net for person re-identification. In *IEEE conference on computer vision and pattern recognition*.

Chen, Y., Zhu, X., & Gong, S. (2017). Person re-identification by deep learning multi-scale representations. In *Workshop of the IEEE international conference on computer vision* (pp. 2590–2600).

Chen, Y., Zhu, X., & Gong, S. (2018a). Deep association learning for unsupervised video person re-identification. In *British machine vision conference*.

Chen, Y., Zhu, X., & Gong, S. (2019). Instance-guided context rendering for cross-domain person re-identification. In *IEEE international conference on computer vision*.

Chen, Y. C., Zhu, X., Zheng, W. S., & Lai, J. H. (2018b). Person re-identification by camera correlation aware feature augmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(2), 392–408.

De Haan, L., & Ferreira, A. (2007). *Extreme value theory: An introduction*. Berlin: Springer.

Deng, W., Zheng, L., Kang, G., Yang, Y., Ye, Q., & Jiao, J. (2018a). Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *IEEE conference on computer vision and pattern recognition*.

Deng, W., Zheng, L., Ye, Q., Kang, G., Yang, Y., & Jiao, J. (2018b). Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person reidentification. In *IEEE conference on computer vision and pattern recognition* (vol. 1, p. 6).

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, *48*(1), 71–99.

Fan, H., Zheng, L., Yan, C., & Yang, Y. (2018). Unsupervised person re-identification: Clustering and fine-tuning. *ACM Transactions on Multimedia Computing, Communications, and Applications*, *14*(4).

Farenzena, M., Bazzani, L., Perina, A., Murino, V., & Cristani, M. (2010). Person re-identification by symmetry-driven accumulation of local features. In *IEEE conference on computer vision and pattern recognition* (pp. 2360–2367). IEEE.

Figueira, D., Bazzani, L., Minh, H. Q., Cristani, M., Bernardino, A., & Murino, V. (2013). Semi-supervised multi-feature learning for person re-identification. In *IEEE international conference on advanced video and signal based surveillance* (pp. 111–116).

Gong, S., Cristani, M., Yan, S., & Loy, C. C. (2014). *Person re-identification*. Berlin: Springer.

Gray, D., & Tao, H. (2008). Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *European conference on computer vision*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Hou, R., Ma, B., Chang, H., Gu, X., Shan, S., & Chen, X. (2019). Interaction-and-aggregation network for person re-identification. In *The IEEE conference on computer vision and pattern recognition*.

Khan, F. M., & Bremond, F. (2016). Unsupervised data association for metric learning in the context of multi-shot person re-identification. In *IEEE international conference on advanced video signal-based surveillance* (pp. 256–262).

Kodirov, E., Xiang, T., & Gong, S. (2015). Dictionary learning with iterative Laplacian regularisation for unsupervised person re-identification. In *British machine vision conference* (p. 8).

Kodirov, E., Xiang, T., Fu, Z., & Gong, S. (2016). Person re-identification by unsupervised l_1 graph learning. In *European conference on computer vision* (pp. 178–195).

Krueger, K. A., & Dayan, P. (2009). Flexible shaping: How learning in small steps helps. *Cognition*, *110*(3), 380–394.

Li, M., Zhu, X., & Gong, S. (2018a). Unsupervised person re-identification by deep learning Tracklet association. In *European conference on computer vision* (pp. 737–753).

Li, M., Zhu, X., & Gong, S. (2019). Unsupervised tracklet person re-identification. In *IEEE transactions on pattern analysis and machine intelligence*.

Li, W., Zhao, R., Xiao, T., & Wang, X. (2014). Deepreid: Deep filter pairing neural network for person re-identification. In *IEEE conference on computer vision and pattern recognition* (pp. 152–159).

Li, W., Zhu, X., & Gong, S. (2017). Person re-identification by deep joint learning of multi-loss classification. In *Proceedings of international joint conference on artificial intelligence*.

Li, W., Zhu, X., & Gong, S. (2018b). Harmonious attention network for person re-identification. In *IEEE conference on computer vision and pattern recognition* (pp. 2285–2294).

Liao, S., Hu, Y., Zhu, X., & Li, S. Z. (2015) Person re-identification by local maximal occurrence representation and metric learning. In *IEEE conference on computer vision and pattern recognition* (pp. 2197–2206).

Lin, S., Li, H., Li, C. T., & Kot, A. C. (2018). Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification. In *British machine vision conference*.

Lin, Y., Dong, A., Zheng, L., Yan, Y., & Yang, Y. (2019). A bottom-up clustering approach to unsupervised person re-identification. In *AAAI conference on artificial intelligence* (vol. 2).

Lisanti, G., Masi, I., Bagdanov, A. D., & Del Bimbo, A. (2014). Person re-identification by iterative re-weighted sparse ranking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*(8), 1629–1642.

Liu, X., Song, M., Tao, D., Zhou, X., Chen, C., & Bu, J. (2014). Semi-supervised coupled dictionary learning for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3550–3557)

Liu, Z., Wang, D., & Lu, H. (2017). Stepwise metric promotion for unsupervised video person re-identification. In *IEEE international conference on computer vision*

Ma, X., Zhu, X., Gong, S., Xie, X., Hu, J., Lam, K. M., et al. (2017). Person re-identification by unsupervised video matching. *Pattern Recognition*, *65*, 197–210.

Maaten, L. V. D., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(11), 2579–2605.

Matsukawa, T., Okabe, T., Suzuki, E., & Sato, Y. (2016). Hierarchical gaussian descriptor for person re-identification. In *IEEE conference on computer vision and pattern recognition* (pp. 1363–1372).

Meng, J., Wu, S., & Zheng, W. (2019). Weakly supervised person re-identification. In *IEEE international conference on computer vision and pattern recognition*. IEEE.

Peng, P., Tian, Y., Xiang, T., Wang, Y., Pontil, M., & Huang, T. (2018). Joint semantic and latent attribute modelling for cross-class transfer learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(7), 1625–1638.

Peng, P., Xiang, T., Wang, Y., Pontil, M., Gong, S., Huang, T., et al. (2016). Unsupervised cross-dataset transfer learning for person re-identification. In *IEEE conference on computer vision and pattern recognition*.

Prosser, B. J., Zheng, W. S., Gong, S., & Xiang, T. (2010). Person re-identification by support vector ranking. In *British machine vision conference* (vol. 2, p. 6).

Quan, R., Dong, X., Wu, Y., Zhu, L., & Yang, Y. (2019). Auto-reid: Searching for a part-aware convnet for person re-identification. In *IEEE international conference on computer vision*.

Ristani, E., Solera, F., Zou, R., Cucchiara, R., & Tomasi, C. (2016). Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV workshop on benchmarking multi-target tracking*.

Scheirer, W. J., de Rezende Rocha, A., Sapkota, A., & Boult, T. E. (2013). Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(7), 1757–1772.

Scheirer, W. J., Jain, L. P., & Boult, T. E. (2014). Probability models for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(11), 2317–2324.

Shen, Y., Li, H., Xiao, T., Yi, S., Chen, D., & Wang, X. (2018a). Deep group-shuffling random walk for person re-identification. In *IEEE conference on computer vision and pattern recognition*.

Shen, Y., Li, H., Yi, S., Chen, D., & Wang, X. (2018b). Person re-identification with deep similarity-guided graph neural network. In *European conference on computer vision*.

Song, C., Huang, Y., Ouyang, W., & Wang, L. (2018). Mask-guided contrastive attention model for person re-identification. In *IEEE conference on computer vision and pattern recognition*.

Sun, Y., Zheng, L., Yang, Y., Tian, Q., & Wang, S. (2018). Beyond part models: person retrieval with refined part pooling (and a strong convolutional baseline). In *European conference on computer vision*.

Tsoumakas, G., & Katakis, I. (2007). Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, *3*(3), 1–13.

Wang, H., Gong, S., & Xiang, T. (2014a). Unsupervised learning of generative topic saliency for person re-identification. In *British machine vision conference*.

Wang, H., Zhu, X., Xiang, T., & Gong, S. (2016a). Towards unsupervised open-set person re-identification. In *IEEE international conference on image processing* (pp. 769–773). IEEE.

Wang, J., Zhu, X., Gong, S., & Li, W. (2018). Transferable joint attribute-identity deep learning for unsupervised person re-

identification. In *IEEE conference on computer vision and pattern recognition* (pp. 2275–2284).

Wang, T., Gong, S., Zhu, X., & Wang, S. (2014b). Person re-identification by video ranking. In *European conference on computer vision* (pp. 688–703). Berlin: Springer.

Wang, T., Gong, S., Zhu, X., & Wang, S. (2016b). Person re-identification by discriminative selection in video ranking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(12), 2501–2514.

Wei, L., Zhang, S., Gao, W., & Tian, Q. (2018). Person transfer GAN to bridge domain gap for person re-identification. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 79–88).

Wu, A., Zheng, W. S., Guo, X., & Lai, J. H. (2019). Distilled person re-identification: Towards a more scalable system. In *The IEEE conference on computer vision and pattern recognition (CVPR)*.

Wu, G., Zhu, X., & Gong, S. (2020). Tracklet self-supervised learning for unsupervised person re-identification. In *AAAI conference on artificial intelligence*.

Xin, X., Wang, J., Xie, R., Zhou, S., Huang, W., & Zheng, N. (2019). Semi-supervised person re-identification using multi-view clustering. *Pattern Recognition*, *88*, 285–297.

Ye, M., Ma, A. J., Zheng, L., Li, J., & Yuen, P. C. (2017). Dynamic label graph matching for unsupervised video re-identification. In *IEEE international conference on computer vision*.

Yu, H. X., Wu, A., & Zheng, W. S. (2017). Cross-view asymmetric metric learning for unsupervised person re-identification. In *IEEE international conference on computer vision* (pp. 994–1002).

Yu, H. X., Wu, A., & Zheng, W. S. (2019a). Unsupervised person re-identification by deep asymmetric metric embedding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yu, H. X., Zheng, W. S., Wu, A., Guo, X., Gong, S., & Lai, J. H. (2019b). Unsupervised person re-identification by soft multilabel learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2148–2157).

Zagoruyko, S., & Komodakis, N. (2016). Wide residual networks. arXiv preprint arXiv:160507146.

Zhang, Z., Lan, C., Zeng, W., & Chen, Z. (2019). Densely semantically aligned person re-identification. In *The IEEE conference on computer vision and pattern recognition (CVPR)*.

Zhao, H., Tian, M., Sun, S., Shao, J., Yan, J., Yi, S., et al. (2017). Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *IEEE conference on computer vision and pattern recognition*.

Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., & Tian, Q. (2015). Scalable person re-identification: A benchmark. In *IEEE international conference on computer vision* (pp. 1116–1124).

Zheng, W. S., Gong, S., & Xiang, T. (2013). Re-identification by relative distance comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(3), 653–668.

Zheng, Z., Zheng, L., & Yang, Y. (2017). Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. In *IEEE conference on computer vision and pattern recognition* (pp. 3754–3762).

Zheng, Z., Yang, X., Yu, Z., Zheng, L., Yang, Y., Kautz, J. (2019). Joint discriminative and generative learning for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2138–2147).

Zhong, Z., Zheng, L., Li, S., & Yang, Y. (2018). Generalizing a person retrieval model hetero- and homogeneously. In *European conference on computer vision*.

Zhong, Z., Zheng, L., Luo, Z., Li, S., & Yang, Y. (2019). Invariance matters: Exemplar memory for domain adaptive person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 598–607).

Zhou, K., Yang, Y., Cavallaro, A., & Xiang, T. (2019). Omni-scale feature learning for person re-identification. In *IEEE international conference on computer vision*.

Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE international conference on computer vision* (pp. 2223–2232).

Zhu, X., Morerio, P., & Murino, V. (2019a). Unsupervised domain adaptive person re-identification based on pedestrian attributes. In *Proceedings of the IEEE international conference on image processing*.

Zhu, X., Zhu, X., Li, M., Murino, V., & Gong, S. (2019b). Intra-camera supervised person re-identification: A new benchmark. In *Workshop of IEEE international conference on computer vision*.