Intra-Camera Supervised Person Re-Identification: A New Benchmark

Xiangping Zhu[1], Xiatian Zhu[2], Minxian Li[3], Vittorio Murino[1,4] and Shaogang Gong[3]

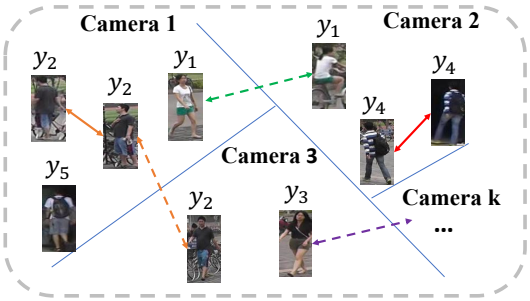
[1]Pattern Analysis and Computer Vision (PAVIS), Istituto Italiano di Tecnologia, [2]Vision Semantics Ltd.
[3]Queen Mary University of London, [4] Department of Computer Science, University of Verona
{xiangping.zhu2010, eddy.zhuxt}@gmail.com, vittorio.murino@iit.it, {m.li, s.gong}@qmul.ac.uk

Abstract

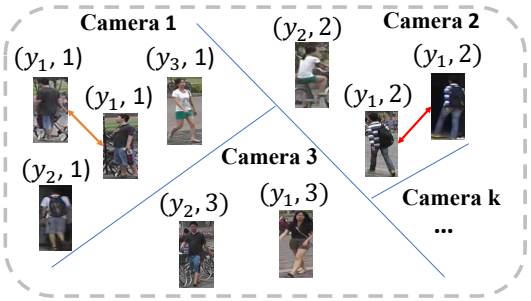
Existing person re-identification (re-id) methods rely mostly on a large set of inter-camera identity labelled training data, requiring a tedious data collection and annotation process therefore leading to poor scalability in practical re-id applications. To overcome this fundamental limitation, we consider person re-identification without *inter-camera identity association but only with identity labels* independently annotated *within each individual camera-view. This eliminates the most time-consuming and tedious inter-camera identity labelling process in order to significantly reduce the amount of human efforts required during annotation. It hence gives rise to a more scalable and more feasible learning scenario, which we call* Intra-Camera Supervised (ICS) *person re-id. Under this ICS setting with weaker label supervision, we formulate a Multi-Task Multi-Label (MTML) deep learning method. Given no inter-camera association, MTML is specially designed for self-discovering the inter-camera identity correspondence. This is achieved by inter-camera multi-label learning under a joint multi-task inference framework. In addition, MTML can also efficiently learn the discriminative re-id feature representations by fully using the available identity labels within each camera-view. Extensive experiments demonstrate the performance superiority of our MTML model over the state-of-the-art alternative methods on three large-scale person re-id datasets in the proposed intra-camera supervised learning setting.*

1. Introduction

Person re-identification (re-id) is a task of reasoning the subtle identity class information in detected person bounding box images captured under non-overlapping camera views [10, 27, 43, 42, 18, 9, 48]. This is still a rather challenging task due to the nonrigid structure of the human body, the highly variable illumination conditions, and low resolution person bounding box images. Most existing deep learning re-id methods in the literature train convolutional neural network (ConvNet) models in a supervised learning fashion [15, 36, 3, 40, 17, 2, 28, 31]. One of the major limitations with supervised modelling is rooted in assuming



(a) Fully supervised re-id data annotation



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

Figure 1. Illustration of person re-id training data annotation: (a) Fully supervised re-id data annotation with a single inter-camera identity label space. y_i denotes a person ID label across cameras; (b) The proposed intra-camera supervised re-id data annotation only with multiple camera-specific identity label spaces. (y_i, p) denotes a person ID y_i with the associated camera index p, i.e., per-camera independent labelling, and here, the same identity label y_i in different camera-views most probably refers to different person. The solid and dashed arrows denote the intra-camera and inter-camera annotations, respectively. Identity is color coded. For simplicity, we use one image in both (a) and (b) to denote one identity. Compared with fully supervised re-id labeling, the inter-camera association labels are not annotated in the proposed re-id data labeling.

the availability of a large set of inter-camera labelled training identity classes collected through an exhaustive and expensive annotation process. This dramatically degrades the usability and scalability of such methods in real-world application and deployments at scales.

This problem has received a significant amount of attentions recently. One intuitive approach is unsupervised person re-id. Existing methods of this kind can be generally divided into three categories: (1) Domain generic feature design [10, 8, 42, 18, 24]; (2) Unsupervised domain adaptation [26, 6, 34, 19, 38, 39, 46, 49, 25]; (3) Unsupervised learning [33, 4, 13, 20, 14]. By hand-crafting universal person appearance features, the first category aims to improve the re-id model performance generically. However, the methods in this category often yield dramatically inferior model generalisation capability due to limited information involved in such representations. The second category attempts to transfer the identity knowledge of a labelled source domain to an unlabelled target domain via image or feature adaptation. Unfortunately, such methods implicitly assume that the source and target domains have reasonably similar camera viewing conditions for ensuring sufficient transferable knowledge. As a more scalable approach, the third category instead leverages only unlabelled target domain data during model training. To benefit from existing supervised learning algorithms, previous unsupervised re-id methods often turn to the idea of self-discovering the underlying identity label information [4, 14, 20]. Compared with conventional manual labelling in supervised methods, this automatical label annotation remains less accurate and less complete, leading to the inferior re-id model optimization and thus much lower re-id performance.

In this work, we instead investigate the person re-id scalability from the data annotation perspective. We consider a more scalable re-id problem with cheaper training data labelling where person identity (ID) labels are annotated in each camera-view *without* inter-camera association. This is based on our observation that inter-camera search in the manual annotation is the most time-consuming and expensive sub-process (Fig 1a). It is because, the generic (unframed) people usually takes a-prior unknown routines in open public space with complex space-time topology. On the other hand, labelling person identity classes in each camera view *independently* is much simpler and faster, possibly further benefiting from off-the-shelf tracking algorithms in a single camera view (Fig 1b). We name this new setting as *Intra-Camera Supervised* (ICS) person re-id. Compared with conventional strong re-id supervision with labelled inter-camera identity association, this re-id problem focuses a learning algorithm on self-discovering the correspondence relationships between camera-specific identity spaces. It presents a new modelling challenge.

To address the ICS re-id problem, we proposed a Multi-Task Multi-Label (MTML) deep learning model in this work. Since there is no inter-camera association in the proposed re-id data labeling, MTML is specially designed for self-discovering the inter-camera identity correspondence by the inter-camera multi-label learning component under a joint multi-task inference framework. Some previous inter-camera identity association re-id methods [4, 13, 14] learn discriminative representations by associating similar samples in the feature space. In contrast, we introduce an idea of multiple labels for each person identity in inter-camera association across different label spaces for better exploiting the per-camera identity labelling information. In addition, MTML can also efficiently learn the discriminative re-id features using the provided per-camera identity labels based on multi-camera multi-task learning.

The **contributions** of this work are: **(1)** We reformulate supervised learning person re-id by removing explicitly the assumption for exhaustive inter-camera pairwise labelling in model training. This eliminates the most time-consuming and tedious inter-camera pairwise ID labelling task required in the conventional re-id model training. In return, a more challenging re-id model learning problem is presented with only per-camera independently annotated ID labels in the training dataset. Compared to completely unsupervised re-id, our introduced re-id problem enables the re-id model to benefit from per-camera view labelled ID information which can be easily annotated or generated using tracking algorithms. **(2)** We formulate a Multi-Task Multi-Label (MTML) learning method for ICS person re-id. MTML takes a multi-task learning framework to jointly account for independent camera-specific identity discriminative labelling information and self-discovering inter-camera identity association in a multi-labelling fashion. Three large-scale re-id datasets, i.e., Market-1501 [42], DukeMTMC-reID [44, 29], and MSMT17 [35], have been used in experiments with the proposed ICS setting. The results demonstrate the superiority of the proposed MTML method compared with the state-of-the-art person re-id models.

2. Related Work

As we are concerned with the re-id scalability issue from the person re-id dataset annotation perspective, this section will discuss and review supervised and unsupervised person re-id works in the literature.

Supervised learning based person re-id methods dominate the literature [36, 35, 3, 32, 41, 50, 5, 16, 2, 40, 30, 17, 28, 31]. This type of models are trained in a *strongly* supervised manner by inter-camera pairwise ID labelled training images. They suffer from significant model performance degradation when the test domain is dissimilar to the training domain. Moreover, supervised learning based re-id models are effective only when strongly labelled training data are available at large scale for every target do-

main. This limits their usefulness. Semi-supervised learning methods [22, 33] decrease the amount of labelled training data but still require some cross-view pairwise labelling. Removing the expensive inter-camera pairwise labelling requirement for re-id model learning is desirable in practice.

Unsupervised learning based person re-id models have received increasing attention with three flavours, i.e., domain generic feature design [10, 8, 42, 18, 24], unsupervised domain adaptation [34, 19, 45, 38, 46] and unsupervised model learning [21, 33, 12, 23, 4, 13, 20, 14]. All these methods do not need labelled training data from the target domain therefore more deployable. However, their re-id performances are much weaker than those of supervised learning based models (when training and test domains are similar).

Intra-camera supervised learning based person re-id is considered in this work, where the strong inter-camera identity association labels are removed from the training data. Without the need for manually annotating identity correspondences between every pair of camera views, we minimise the amount of labours required for person identity class annotation and enable a re-id model to be more feasible in deployment at scale. To solve this re-id problem, we develop a re-id learning algorithm for making full use of per-camera independently annotated ID labels and self-discovering most likely person correspondences between different camera views, yielding stronger re-id models than the unsupervised learning counterpart.

3. Methodology

In this section we formulate a person re-id learning method without inter-camera identity association in the training data. Suppose there are M camera views in a camera network. For the p-th camera view, we *independently* annotate a set of samples $\mathcal{D}_p = \{(\mathbf{x}_i, y_k, p)\}$ where $\mathbf{x}_i$ is the i-th person image in $\mathcal{D}_p$. Each person image $\mathbf{x}_i$ is associated with an identity label $y_k \in \{y_1, y_2, \cdots, y_{N_p}\}$ and the corresponding camera identity $p \in \{1, 2, \cdots, M\}$. N_p is the total number of unique person identities in $\mathcal{D}_p$. Due to per-camera independent labelling nature, the same identity labels (e.g. identity y_1) of any two camera views are very likely referring to two different persons. By combining camera-specific labelled data, we obtain the entire training set as $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_M\}$. The presence of such multiple identity class label spaces prevents the training of a conventional supervised re-id model and thus a new effective re-id method is needed.

We formulate a Multi-Task Multi-Label (MTML) deep learning method for addressing this more challenging and more scalable ICS re-id problem. Given the per-camera identity independent labelling nature, the key of model learning lies in two aspects: (1) How to effectively exploit the per-camera identity labels, and (2) How to associate the identity label classes across camera views (or label spaces). MTML achieves these two aspects by integrating two corresponding components: (i) Multi-camera multi-task learning that assigns a separate learning task to each individual camera view for modelling the respective identity space, (ii) inter-camera Multi-label learning that automatically self-discovers the identity associations across camera views and assigns multiple labels to these associated identities for modeling the inter-camera person appearance variation. More details about these two components are presented in the following parts and Fig. 2 gives an overview of the proposed MTML model.

3.1. Multi-Camera Multi-Task Learning

As shown in Fig. 2b and 2c, we consider the multi-task learning strategy [1] given camera independently labelled person identity information. This aims at better mining the common knowledge shared across camera views whilst enhancing model learning by augmented training data for each camera view. Each camera view is treated separately due to their independent labelling property. Importantly, this also allows to derive a person re-id representation with implicit inter-camera identity discriminative capability for facilitating inter-camera identity association [14].

Formally, we create a camera-shared feature representation upon which multi-task branches are rooted. Each branch is responsible for the classification task in a specific camera view. During model training, each task branch can be used to propagate the respective per-camera identity label information via the softmax cross-entropy loss function. For one sample $(\mathbf{x}_i, y_k, p) \in \mathcal{D}_p$, its corresponding softmax cross-entropy loss function can be formulated as:

$$\mathcal{L}^p_{\text{mt},i} = -\mathbb{1}(y_k)\log f_p(\mathbf{v}_i) \tag{1}$$

where $\mathbf{v}_i \in \mathbb{R}^{d\times 1}$ specifies the *camera-shared* feature vector of the corresponding image $\mathbf{x}_i$ from the p-th camera and it is extracted after the fully connected layer FC-d as shown in Fig. 2, in which d is the dimension of a feature vector. $f_p(\cdot) : \mathbb{R}^{d\times 1} \rightarrow \mathbb{R}^{N_p\times 1}$ denotes the classifier function of camera p. The one-hot encoding function $\mathbb{1}(\cdot) : \mathbb{R} \rightarrow \mathbb{R}^{1\times N_p}$ returns an one-hot vector with value 1 for the element at the given index. For stochastic mini-batch deep learning, the *multi-camera multi-task learning* (MT) objective is designed as:

$$\mathcal{L}_{\text{mt}} = \frac{1}{B}\sum_{p=1}^{M}\mathcal{L}^p_{\text{mt}} \tag{2}$$

where $\mathcal{L}^p_{\text{mt}}$ denotes the accumulated cross-entropy loss (Eq. (1)) of all in-batch images from the p-th camera, and B is the mini-batch size. With this multi-camera multi-task learning, the discriminative re-id features can be efficiently learned using the existing identity labels within each camera-view.

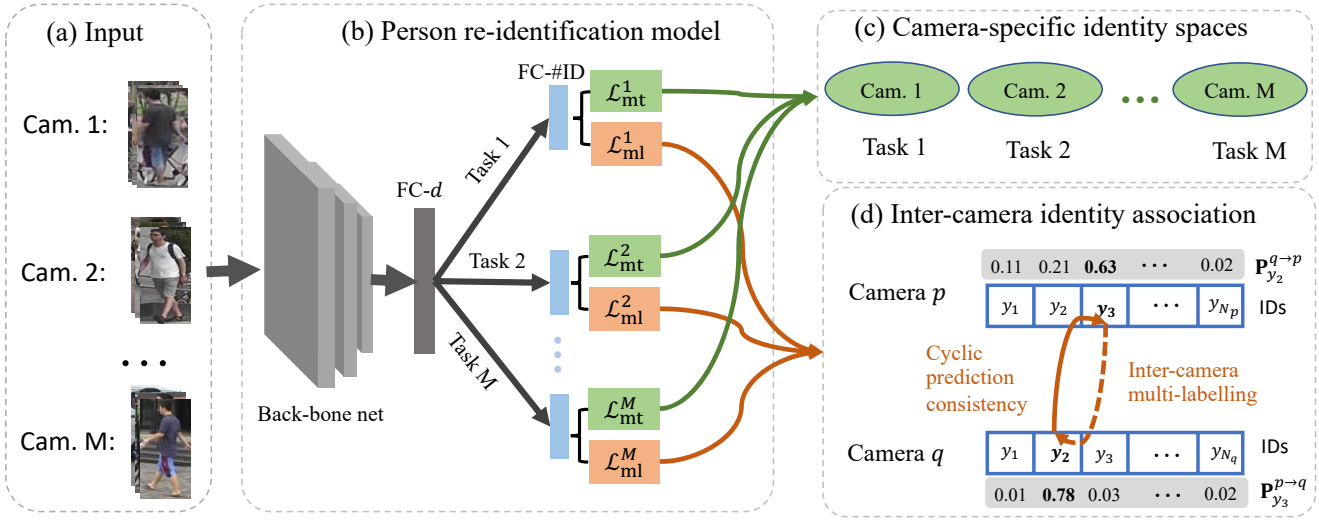


Figure 2. Overview of the proposed Multi-Task Multi-Label (MTML) deep learning method for intra-camera supervised person re-id. **(a)** Given input person images which are labeled independently in each camera view, MTML aims to derive **(b)** an identity discriminative feature representation. This is achieved by designing two components: (1) *multi-camera multi-task learning* where each individual camera view is assigned with **(c)** a separate supervised learning task for modelling the corresponding identity space, and (2) *inter-camera multi-label learning* where **(d)** inter-camera identity association labels are self-discovered using the cyclic prediction consistency strategy. These association labels are further included into the training dataset to impose the re-id model to modeling the the inter-camera person appearance variation under the multi-task inference framework. The re-id model is composed with the backbone network and fully connected layers, i.e., FC-d and FC-#ID, in which d denotes the dimension of the feature representation and #ID denotes the number of identities in corresponding camera-view.

3.2. Inter-Camera Multi-Label Learning

In person re-id, inter-camera person appearance variation is one of the most significant elements during model training. Whilst this is implicitly learned in the multi-camera multi-task learning component as discussed above, it is insufficient to fully capture the underlying inter-camera identity correspondence relationships. To address this problem, an inter-camera multi-label learning component is designed that aims to self-discover the identity correspondence between camera-specific identity label spaces and imposes the re-id model to effectively model the inter-camera person appearance variation.

Specifically, given an identity class $y_k \in \{y_1, y_2, \cdots, y_{N_p}\}$ from camera $p \in \{1, 2, \ldots, M\}$, we want to find if a true match exists in camera q. To this end, all the person images of y_i are mapped into the branch of camera q and an average prediction of y_i in camera q is obtained as:

$$\mathbf{P}_{y_k}^{p \to q} = \text{avg}(f_q(\mathbf{v}_i)), \tag{3}$$

in which $\mathbf{P}_{y_k}^{p \to q} \in \mathbb{R}^{N_q \times 1}$ is the averaged prediction of y_k in camera q. As in Eq. (1), $f_q(\cdot)$ denotes the classifier function of camera q and avg$(\cdot)$ is the averaging function. We then nominate the identity class in camera q with the maximum likelihood probability as the candidate matching identity:

$$l^* = \arg \max_{l \in \{1,2,\ldots,N_q\}} \mathbf{P}_{y_k}^{p \to q}(l), \tag{4}$$

where l^* is the index of identity $y_{l^*} \in \{y_1, y_2, \cdots, y_{N_q}\}$ in q-th camera. To boost the accuracy and robustness of matching pairs, the identity y_{l^*} is further mapped back to camera p in a similar way as Eq. (3) and the corresponding candidate matching identity y_{t^*} as in Eq. (4) is retrieved. This cyclic mapping and matching operation between every two camera views determines the inter-camera identity association result as:

$$\begin{cases} (y_k, y_{l^*}) \text{ is a matching pair,} & \text{if } y_{t^*} = y_k, \\ (y_k, y_{l^*}) \text{ is a unmatching pair,} & \text{otherwise.} \end{cases} \tag{5}$$

To benefit model training from self-discovered identity matching pairs, a proper supervision function is designed. Considering the idea of inter-camera prediction based identity association, the inter-camera learning is performed by multi-labeling the associated person identities.

In particular, given an identity matching pair (y_k, y_{l^*}), with y_k from camera p and y_{l^*} from camera q as defined in Eq. (5), the person images corresponding to the associated matching identities y_k and y_{l^*} are assigned with multiple labels. For the person images of y_k, they are assigned with the new label (y_{l^*}, q) and similarly, the person images of y_{l^*} are assigned with the new label (y_k, p). After this inter-camera multi-labeling, the images of y_k and y_{l^*} are attached with the identical multiple identity labels, i.e., (y_k, p) and (y_{l^*}, q), and thus y_k and y_{l^*} are inter-camera associated.

With these newly assigned labels, a simple but efficient

loss function is designed based on the softmax cross entropy loss as the MT loss in Eqs. (1) and (2). For one person image $\mathbf{x}_i$ with new label (y_{l^*}, q) as in Eq. (4), its ML loss can be formulated as:

$$\mathcal{L}^q_{\text{ml},i} = -\mathbb{1}(y_{l^*})\log f_q(\mathbf{v}_i) \tag{6}$$

in which $\mathbf{v}_i$ is the feature vector corresponding to the person image $\mathbf{x}_i$. The definitions of $\mathbb{1}(\cdot)$ and $f_q(\cdot)$ are the same as in Eq. (1). As the MT loss in Eq. (2), the final ML loss for one mini-batch is defined as:

$$\mathcal{L}_{\text{ml}} = \frac{1}{M}\sum_{q=1}^{M}\mathcal{L}^q_{\text{ml}} \tag{7}$$

in which $\mathcal{L}^q_{\text{ml}}$ is the ML loss in camera q, i.e., $\mathcal{L}^q_{\text{ml}} = \frac{1}{b_q}\sum_{i=1}^{b_q}\mathcal{L}^q_{\text{ml},i}$. b_q is the number of images with newly assigned labels in camera q.

3.3. Model Objective Loss Function

By combining the multi-camera and multi-label learning function in a multi-task manner, we obtain the final MTML model objective loss function as:

$$\mathcal{L} = \mathcal{L}_{\text{mt}} + \lambda\mathcal{L}_{\text{ml}}, \tag{8}$$

where λ controls the relative weight of the two terms. In our experiments, we set $\lambda = 0.5$ considering that inter-camera identity association is necessarily noisy therefore adversely affects the quality of $\mathcal{L}_{\text{ml}}$.

3.4. Model Training and Inference

The stochastic gradient descent algorithm can be applied for optimizing the proposed deep re-id model. In the considered re-id dataset annotation as shown in Fig. 1(b), per-camera identity labels are accurately annotated whilst the identity matching between camera views is likely inaccurate. Based on this observation, the proposed re-id model is first pre-trained using only the multi-camera multi-task learning loss $\mathcal{L}_{\text{mt}}$ (Eq (2)). Then based on this pre-trained re-id model, the inter-camera multi-label learning is iteratively performed using the model objective loss function $\mathcal{L}$ as in Eq. (8). In every iteration, the re-id model will be first trained for a number of epochs, and then the cyclic prediction consistency and inter-camera multi-labeling will be applied for associating inter-camera identities. The newly assigned multi-labels of associated identities will be further included into the training dataset for model learning in the following iteration.

In model inference, the trained MTML model is deployed to extract the camera-shared features of test person images as their re-id representations. For efficient re-id matching and ranking, the Euclidean distance metric is utilised to compute the probe-gallery pairwise similarity.

Table 1. Comparisons between the proposed MTML method and existing re-id methods on Market1501 dataset.

Metric	R1	R10	R20	mAP
PUL [7]	41.9	64.3	70.5	18.0
CAMEL [37]	54.5	-	-	26.3
TJ-AIDL [34]	58.2	-	-	26.5
CycleGAN [47]	48.1	-	-	20.7
SPGAN [6]	51.5	76.8	82.4	22.8
SPGAN+LMP [6]	58.1	82.7	87.9	26.9
HHL [45]	62.2	84.0	88.3	31.4
MAR [39]	67.7	-	-	40.0
ECN [46]	75.1	91.6	-	43.0
E-PCSL	42.6	64.6	69.9	17.6
UTAL [14]	69.2	85.5	89.7	46.2
MTML	**85.3**	**96.2**	**97.6**	**65.2**

4. Experiment

4.1. Experimental Setup

Datasets. Three large-scale re-id datasets, i.e., Market-1501 [42], DukeMTMC-reID [44, 29], and MSMT17 [35], are selected for evaluating our proposed ICS problem and our MTML method. As no existing re-id datasets annotated in the ICS fashion, we adopted these three fully labelled re-id datasets by independently annotating their identity labels in each camera-view as shown in Fig. 1. We still utilise the identical test data of each dataset for model performance evaluation. We will publicly release these ICS person re-id benchmarks.

Performance metrics. We used the common Cumulative Matching Characteristic (CMC) and mean Average Precision (mAP) metrics for model performance measurement.

Implementation details. In practice, the ImageNet pre-trained ResNet-50 [11] is selected as the backbone CNN [2] of our MTML model. For multi-task learning, each branch is formed by a FC classification layer. Person bounding box images are resized to 256×128 in pixel before feeding into the network. The standard stochastic gradient descent (SGD) optimizer is adopted for training the MTMC model with the initial learning rate of 0.05. In pre-training the model using only the MT loss, the learning rate is decayed 10 times every 40 epochs and the epoch number is set to 100. In inter-camera multi-label learning, the learning rate is decayed 10 times after 8 epochs. The epoch number is set to 15 in each iteration and the number of iteration is set to 8. In order to balance the model training speed across camera views, we randomly selected from each camera view the same number of images (i.e., 4 images) from one person identity and the same number of persons (i.e., 2 persons). By default, we set $\lambda = 0.5$ in Eq (8) for balancing

[2]Layers after avg-pooling are removed.

Table 2. Comparisons between the proposed MTML method and existing re-id methods on the DukeMTMC-reID dataset.

Metric	R1	R10	R20	mAP
PUL [7]	23.0	39.5	44.2	12.0
TJ-AIDL [34]	44.3	-	-	23.0
CycleGAN [47]	38.5	-	-	19.9
SPGAN [6]	41.1	63.0	69.6	22.3
SPGAN+LMP [6]	46.9	68.5	74.0	26.4
HHL [45]	46.9	66.7	71.9	27.2
MAR [39]	67.1	-	-	48.0
ECN [46]	63.3	80.4	-	40.4
E-PCSL	38.8	58.9	64.6	22.1
UTAL [14]	62.3	80.7	84.4	44.6
MTML	**71.7**	**86.9**	**89.6**	**50.7**

Table 3. Comparisons between the proposed MTML method and existing re-id methods on the MSMT17 dataset.

Metric	R1	R10	R20	mAP
PTGAN [35]	11.8	27.4	-	3.3
ECN [46]	30.2	46.8	-	10.2
E-PCSL	16.8	31.5	37.4	5.4
UTAL [14]	31.4	51.0	58.1	13.1
MTML	**44.1**	**63.9**	**70.0**	**18.6**

the the losses of $\mathcal{L}_{\text{mt}}$ and $\mathcal{L}_{\text{ml}}$.

4.2. Evaluation on Person Re-Identification

Evaluated methods. Apart from the proposed MTML model, we further evaluated two methods particularly adapted to the newly introduced ICS person re-id setting: **(1)** *Ensemble of Per-Camera Supervised Learning* (E-PCSL): Without inter-camera ID labels, we trained a separate re-id model for each camera on the corresponding labelled training data. We used the ResNet-50 as the backbone CNN, and the softmax cross-entropy loss function as the supervised objective. During deployment, given a test image we extracted the feature vectors of all the per-camera models, concatenated them into a single representation vector, and utilised the Euclidean distance for re-id matching. **(2)** *Unsupervised Tracklet Association Learning* (UTAL) [14]: This method is designed for associating person tracklets in an unsupervised manner for video based re-id, taking the auto-detected tracklets as imagery data form in particular. For enabling multi-shot image based re-id as considered here, following UTAL we stacked the images with the same ID from the same camera into a single tracklet, forming the intra-camera supervision specifically for this model. In terms of experiment setting, UTAL assumes the same training data annotation as in our work. However, it is noteworthy to mention that this is due to lacking spatial-temporal information in the existing image based person re-id benchmarks; Conceptually, the two works investigate rather different person re-id scenarios, starting from distinctive motivations and annotation assumptions.

In addition, the proposed MTML is also compared with the state-of-the-art unsupervised domain adaptive re-id methods which consider the re-id problem with fully labelled data in the source domain but no labels in the target domain. These methods include CAMEL [37], PUL [7], TJ-AIDL [34], CycleGAN [47], SPGAN [6], PTGAN [35], HHL [45], MAR [39], ECN [46]. This provides a overall quantitative evaluation and comparision between different person re-id settings, but no apple-to-apple comparison due to different types of supervision involved.

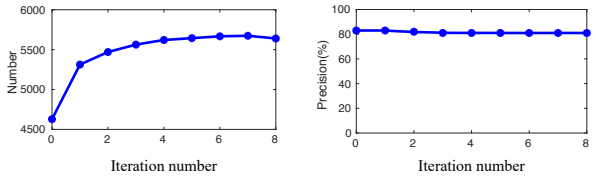
Results. Tables 1-3 give the re-id performance comparison results between our MTML model and other considered methods. Several observations can be derived that: **(1)** By independently exploiting camera-specific identity class annotation, the baseline E-PCSL yields the weakest re-id model generalisation. This is due to the incapability of leveraging the shared knowledge between camera views and mining the inter-camera identity matching information. **(2)** The model performance is continuously increased by more recent unsupervised re-id models. In comparison, the proposed MTML model improves the performance observably. One reason is that our model benefits from more scalable per-camera ID labelling, in addition to the superior formulation of our model. **(3)** The proposed MTML model significantly outperforms both E-PCSL and UTAL, suggesting the performance superiority of our method in tackling the person re-id problem under the proposed cheaper annotation case. Whilst MTML shares partly the model structure with UTAL in terms of multi-task learning design, we observed a large performance difference between them. The plausible reason may be due to the unique advantage of exploiting the cyclic prediction consistency based inter-camera identity association.

Table 4. Effectiveness analysis of two components in MTML: multi-camera multi-task (MT) and inter-camera multi-label (ML).

dataset	Market-1501			
metric	R1	R10	R20	mAP
MT	78.4	93.1	95.7	52.1
MTML	**85.3**	**96.2**	**97.6**	**65.2**
dataset	DukeMTMC-reID			
metric	R1	R10	R20	mAP
MT	65.2	81.1	85.6	44.7
MTML	**71.7**	**86.9**	**89.6**	**50.7**
dataset	MSMT17			
metric	R1	R10	R20	mAP
MT	39.6	59.6	65.7	15.9
MTML	**44.1**	**63.9**	**70.0**	**18.6**

4.3. Further Analysis and Discussions

Model component analysis. We examined the effectiveness of the two model components in MTML, i.e., multi-camera multi-task (MT) and inter-camera multi-label (ML) learning. Table 4 shows that: (1) With the MT component alone, the model can already achieve very competitive performance, suggesting the significance of sharing labelling knowledge across all the camera views via joint multi-task inference. (2) After adding the ML component, the model generalisation capability can be further boosted. This indicates the positive influences of leveraging the inter-camera identity association information through self-supervision despite at the risk of deriving false inter-camera identity associations and propagating their error information into re-id model during training.



(a) Number of matching pairs (b) Precision of matching pairs

Figure 3. Dynamics of self-discovered inter-camera identity matching pairs during model training on the Market-1501 dataset.

Inter-camera identity association dynamics. To further examine the benefits of inter-camera identity association, we tracked the evolving dynamics of self-discovered matching pairs during the model training process. Figure 3 shows that our model is able to reveal an increasingly number of inter-camera identity matching pairs whilst maintaining high association accuracy. This explicitly explains the model performance advantage of the proposed inter-camera multi-label learning idea.

5. Conclusion

In this work, we presented a more scalable intra-camera supervised (ICS) person re-identification problem, characterised by re-id model learning without cross-view pairwise labelling but with only per-camera independent person identity labels. The key idea is to eliminate the tedious process of manually annotating exhaustively identity classes across every pair of camera views in a surveillance network, both costly and sparsely available. This reformulates the conventional supervised re-id model learning into a weakly supervised learning problem with multiple independent ID label spaces across camera views. Consequently, it focuses the learning task on self-discovering inter-camera identity label associations. To that end, we introduced a Multi-Task Multi-Label (MTML) learning algorithm capable of fully exploiting the available weak re-id supervision constraint whilst simultaneously self-mining inter-camera identity association by a cyclic classification consistency idea. Extensive evaluations were conducted on three re-id benchmarks to validate the advantages of the proposed MTML model over the state-of-the-art alternative methods in the proposed ICS learning setting. The detailed component analysis is also provided for giving insights on our model design.

Acknowledgement

This work was partially supported by Vision Semantics Limited, the Alan Turing Institute Fellowship Project on Deep Learning for Large-Scale Video Semantic Search, and the Innovate UK Industrial Challenge Project on Developing and Commercialising Intelligent Video Analytics Solutions for Public Safety (98111-571149).

References

[1] A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. In *Advances in Neural Information Processing Systems*, pages 41–48, 2007. 3

[2] X. Chang, T. M. Hospedales, and T. Xiang. Multi-level factorisation net for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2109–2118, 2018. 1, 2

[3] Y. Chen, X. Zhu, and S. Gong. Person re-identification by deep learning multi-scale representations. In *IEEE International Conference on Computer Vision*, pages 2590–2600, 2017. 1, 2

[4] Y. Chen, X. Zhu, and S. Gong. Deep association learning for unsupervised video person re-identification. In *British Machine Vision Conference*, 2018. 2, 3

[5] Y.-C. Chen, X. Zhu, W.-S. Zheng, and J.-H. Lai. Person re-identification by camera correlation aware feature augmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 2

[6] W. Deng, L. Zheng, G. Kang, Y. Yang, Q. Ye, and J. Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person reidentification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2, 5, 6

[7] H. Fan, L. Zheng, C. Yan, and Y. Yang. Unsupervised person re-identification: Clustering and fine-tuning. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 14(4), 2018. 5, 6

[8] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010. 2, 3

[9] S. Gong, M. Cristani, S. Yan, and C. C. Loy. *Person re-identification*. Springer, 2014. 1

[10] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *European Conference on Computer Vision*, 2008. 1, 2, 3

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 5

[12] E. Kodirov, T. Xiang, Z. Fu, and S. Gong. Person re-identification by unsupervised l1 graph learning. In *European Conference on Computer Vision*, pages 178–195. Springer, 2016. 3

[13] M. Li, X. Zhu, and S. Gong. Unsupervised person re-identification by deep learning tracklet association. In *European Conference on Computer Vision*, pages 737–753, 2018. 2, 3

[14] M. Li, X. Zhu, and S. Gong. Unsupervised tracklet person re-identification. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 2, 3, 5, 6

[15] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 152–159, 2014. 1

[16] W. Li, X. Zhu, and S. Gong. Person re-identification by deep joint learning of multi-loss classification. In *International Joint Conferences on Artificial Intelligence*, 2017. 2

[17] W. Li, X. Zhu, and S. Gong. Harmonious attention network for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1, 2

[18] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2197–2206, 2015. 1, 2, 3

[19] S. Lin, H. Li, C.-T. Li, and A. C. Kot. Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification. In *British Machine Vision Conference*, 2018. 2, 3

[20] Y. Lin, X. Dong, L. Zheng, Y. Yan, and Y. Yang. A bottom-up clustering approach to unsupervised person re-identification. In *AAAI Conference on Artificial Intelligence*, volume 2, 2019. 2, 3

[21] G. Lisanti, I. Masi, A. D. Bagdanov, and A. Del Bimbo. Person re-identification by iterative re-weighted sparse ranking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(8):1629–1642, 2015. 3

[22] X. Liu, M. Song, D. Tao, X. Zhou, C. Chen, and J. Bu. Semi-supervised coupled dictionary learning for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3550–3557, 2014. 3

[23] X. Ma, X. Zhu, S. Gong, X. Xie, J. Hu, K.-M. Lam, and Y. Zhong. Person re-identification by unsupervised video matching. *Pattern Recognition*, 65:197–210, 2017. 3

[24] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato. Hierarchical gaussian descriptor for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1363–1372, 2016. 2, 3

[25] R. Panda, A. Bhuiyan, V. Murino, and A. K. Roy-Chowdhury. Unsupervised adaptive re-identification in open world dynamic camera networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7054–7063, 2017. 2

[26] P. Peng, T. Xiang, Y. Wang, M. Pontil, S. Gong, T. Huang, and Y. Tian. Unsupervised cross-dataset transfer learning for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2

[27] B. J. Prosser, W.-S. Zheng, S. Gong, and T. Xiang. Person re-identification by support vector ranking. In *British Machine Vision Conference*, volume 2, page 6, 2010. 1

[28] X. Qian, Y. Fu, T. Xiang, W. Wang, J. Qiu, Y. Wu, Y.-G. Jiang, and X. Xue. Pose-normalized image generation for person re-identification. In *European Conference on Computer Vision*, 2018. 1, 2

[29] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV workshop on Benchmarking Multi-Target Tracking*, 2016. 2, 5

[30] Y. Shen, H. Li, S. Yi, D. Chen, and X. Wang. Person re-identification with deep similarity-guided graph neural network. In *European Conference on Computer Vision*, September 2018. 2

[31] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *European Conference on Computer Vision*, 2018. 1, 2

[32] H. Wang, X. Zhu, S. Gong, and T. Xiang. Person re-identification in identity regression space. *International journal of computer vision*, 126(12):1288–1310, 2018. 2

[33] H. Wang, X. Zhu, T. Xiang, and S. Gong. Towards unsupervised open-set person re-identification. In *IEEE International Conference on Image Processing*, pages 769–773. IEEE, 2016. 2, 3

[34] J. Wang, X. Zhu, S. Gong, and W. Li. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2275–2284, 2018. 2, 3, 5, 6

[35] L. Wei, S. Zhang, W. Gao, and Q. Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 79–88, 2018. 2, 5, 6

[36] T. Xiao, H. Li, W. Ouyang, and X. Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1249–1258. IEEE, 2016. 1, 2

[37] H.-X. Yu, A. Wu, and W.-S. Zheng. Cross-view asymmetric metric learning for unsupervised person re-identification. In *IEEE International Conference on Computer Vision*, pages 994–1002, 2017. 5, 6

[38] H.-X. Yu, A. Wu, and W.-S. Zheng. Unsupervised person re-identification by deep asymmetric metric embedding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 2, 3

[39] H.-X. Yu, W.-S. Zheng, A. Wu, X. Guo, S. Gong, and J.-H. Lai. Unsupervised person re-identification by soft multilabel learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2148–2157, 2019. 2, 5, 6

[40] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu. Deep mutual learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4320–4328, 2018. 1, 2

[41] L. Zhao, X. Li, Y. Zhuang, and J. Wang. Deeply-learned part-aligned representations for person re-identification. In

IEEE International Conference on Computer Vision, pages 3219–3228, 2017. 2

[42] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *IEEE International Conference on Computer Vision*, pages 1116–1124, 2015. 1, 2, 3, 5

[43] W.-S. Zheng, S. Gong, and T. Xiang. Re-identification by relative distance comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(3):653–668, 2013. 1

[44] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3754–3762, 2017. 2, 5

[45] Z. Zhong, L. Zheng, S. Li, and Y. Yang. Generalizing a person retrieval model hetero- and homogeneously. In *European Conference on Computer Vision*, 2018. 3, 5, 6

[46] Z. Zhong, L. Zheng, Z. Luo, S. Li, and Y. Yang. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 598–607, 2019. 2, 3, 5, 6

[47] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision*, pages 2223–2232, 2017. 5, 6

[48] X. Zhu, A. Bhuiyan, M. L. Mekhalfi, and V. Murino. Exploiting gaussian mixture importance for person re-identification. In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE, 2017. 1

[49] X. Zhu, P. Morerio, and V. Murino. Unsupervised domain adaptive person re-identification based on pedestrian attributes. In *Proceedings of the IEEE International Conference on Image Processing,* In Press. 2

[50] X. Zhu, B. Wu, D. Huang, and W.-S. Zheng. Fast open-world person re-identification. *IEEE Transactions on Image Processing*, 27(5):2286–2300, 2017. 2