Unsupervised Person Re-identification by Deep Learning Tracklet Association

Minxian Li[1,2(✉)], Xiatian Zhu[3], and Shaogang Gong[2]

[1] Nanjing University of Science and Technology, Nanjing, China
minxianli@njust.edu.cn
[2] Queen Mary University of London, London, UK
s.gong@qmul.ac.uk
[3] Vision Semantics Limited, London, UK
eddy@visionsemantics.com

Abstract. Most existing person re-identification (re-id) methods rely on *supervised* model learning on per-camera-pair *manually* labelled pairwise training data. This leads to poor scalability in practical re-id deployment due to the lack of exhaustive identity labelling of image positive and negative pairs for every camera pair. In this work, we address this problem by proposing an unsupervised re-id deep learning approach capable of incrementally discovering and exploiting the underlying re-id discriminative information from *automatically* generated person tracklet data from videos in an end-to-end model optimisation. We formulate a *Tracklet Association Unsupervised Deep Learning* (TAUDL) framework characterised by jointly learning per-camera (within-camera) tracklet association (labelling) and cross-camera tracklet correlation by maximising the discovery of most likely tracklet relationships across camera views. Extensive experiments demonstrate the superiority of the proposed TAUDL model over the state-of-the-art unsupervised and domain adaptation re-id methods using six person re-id benchmarking datasets.

Keywords: Person re-identification · Unsupervised learning
Tracklet · Surveillance video

1 Introduction

Person re-identification (re-id) aims to match the underlying identities of person bounding box images detected from non-overlapping camera views [15]. In recent years, extensive research attention has been attracted [1,7,10,11,14,18,29–31,43,45,52,57] to address the re-id problem. Most existing re-id methods, in particular deep learning models, adopt the *supervised* learning approach. These supervised deep models assume the availability of a large number of *manually* labelled *cross-view identity (ID) matching image pairs* for each camera pair in order to induce a feature representation or a distance metric function optimised

V. Ferrari et al. (Eds.): ECCV 2018, LNCS 11208, pp. 772–788, 2018.
https://doi.org/10.1007/978-3-030-01225-0_45

just for that camera pair. This assumption is inherently limited for generalising a re-id model to many different camera networks therefore cannot scale in practical deployments[1].

It is no surprise then that person re-id by *unsupervised* learning has become a focus in recent research where per-camera pairwise ID labelled training data is not required in model learning [22,24,25,32,35,37,46,48,54,58]. However, all these classical unsupervised learning models are significantly weaker in re-id performance than the supervised models. This is because the lack of cross-view pairwise ID labelled data deprives a model's ability to learn from strong context-aware ID discriminative information in order to cope with significant visual appearance change between every camera pair, as defined by a triplet verification loss function. An alternative approach is to leverage jointly (1) unlabelled data from a target domain which is freely available, e.g. videos of thousands of people travelling through a camera view everyday in a public scene; and (2) pairwise ID labelled datasets from independent source domains [13,38,42,49,55]. The main idea is to first learn a "view-invariant" representation from ID labelled source data, then adapt the model to a target domain by using only unlabelled target data. This approach makes an implicit assumption that the source and target domains share some common cross-view characteristics and a view-invariant representation can be estimated, which is not always true.

In this work, we consider a *pure* unsupervised person re-id deep learning problem. That is, no ID labelled training data is assumed, neither cross-view nor within-view ID labelling. Although this learning objective is similar to two domain transfer models [13,49], both those models do require *suitable*, i.e. visually similar to the target domain, person identity labelled source domain training data. Specifically, we consider unsupervised re-id model learning by jointly optimising unlabelled person tracklet data *within-camera* view to be more discriminative and *cross-camera* view to be more associative in an end-to-end manner.

Our **contributions** are: We formulate a novel unsupervised person re-id deep learning method using person tracklets without the need for camera pairwise ID labelled training data, i.e. *unsupervised tracklet re-id discriminative learning*. Specifically, we propose a **Tracklet Association Unsupervised Deep Learning** (TAUDL) model with two key innovations: (1) *Per-Camera Tracklet Discrimination Learning* that optimises "local" within-camera tracklet label discrimination for facilitating cross-camera tracklet association given per-camera independently created tracklet label spaces. (2) *Cross-Camera Tracklet Association Learning* that maximises "global" cross-camera tracklet label association. This is formulated as to maximise jointly cross-camera tracklet similarity and within-camera tracklet dissimilarity in an end-to-end deep learning framework.

Comparative experiments show the advantages of TAUDL over the state-of-the-art unsupervised and domain adaptation person re-id models using six benchmarks including three multi-shot image based and three video based re-

[1] Exhaustive manual ID labelling of person image pairs for every camera-pair is prohibitively expensive as there are a quadratic number of camera pairs in a network.

id datasets: CUHK03 [29], Market-1501 [60], DukeMTMC [41], iLIDS-VID [50], PRID2011 [19], and MARS [59].

2 Related Work

Most existing re-id models are built by *supervised* model learning on a separate set of per-camera-pair ID labelled training data [1,7–11,18,20,29–31,43,45,47,51,52,57,62]. Hence, their scalability and usability is poor for real-world re-id deployments where no such large training sets are available for every camera pair. Classical unsupervised learning methods based on hand-crafted features offer poor re-id performance [14,22,24,25,32,35,37,46,48,54,58] when compared to the supervised learning based re-id models. While a balancing trade-off between model scalability and re-id accuracy can be achieved by semi-supervised learning [33,48], these models still assume sufficiently large sized cross-view pairwise labelled data for model training. More recently, there are some attempts on unsupervised learning of domain adaptation models [13,38,42,49,55]. The main idea is to explore knowledge from pairwise labelled data in "related" source domains with model adaptation on unlabelled target domain data. Whilst these domain adaptation models perform better than the classical unsupervised learning methods (Tables 2 and 3), they requires implicitly similar data distributions and viewing conditions between the labelled source domain and the unlabelled target domains. This restricts their scalability to arbitrarily diverse (and unknown) target domains.

In contrast to all these existing unsupervised learning re-id methods, the proposed tracklet association based method enables unsupervised re-id deep end-to-end learning from scratch without any assumption on either the scene characteristic similarity between source and target domains, or the complexity of handling identity label space (or lack of) knowledge transfer in model optimisation. Instead, our method directly learns to discover the re-id discriminative knowledge from *unsupervised* tracklet label data automatically generated and annotated from the video data using a common deep learning network architecture. Moreover, this method does not assume any overlap of person ID classes across camera views, therefore scalable to any camera networks without any knowledge about camera space-time topology and/or time-profiling on people cross-view appearing patterns [36]. Compared to classical unsupervised methods relying on extra hand-crafted features, our method learns tracklet based re-id discriminative features from an end-to-end deep learning process. To our best knowledge, this is the *first* attempt at unsupervised tracklet association based person re-id deep learning model without relying on any ID labelled training data (either videos or images).

3 Unsupervised Deep Learning Tracklet Association

To overcome the limitation of *supervised re-id model training*, we propose a novel **Tracklet Association Unsupervised Deep Learning** (TAUDL) approach to

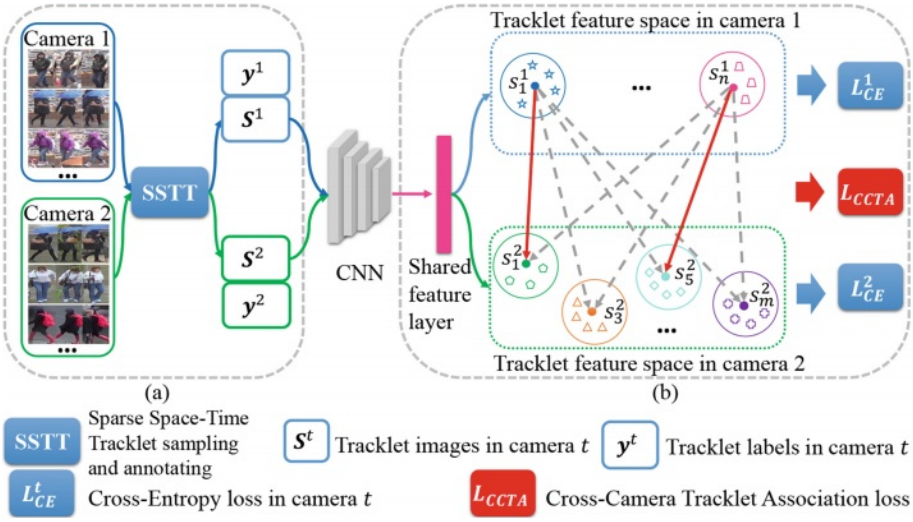

Fig. 1. An overview of Tracklet Association Unsupervised Deep Learning (TAUDL) re-id model: (a) Per-camera unsupervised tracklet sampling and label assignment; (b) Joint learning of both within-camera tracklet discrimination and cross-camera tracklet association in an end-to-end global deep learning on tracklets from all the cameras.

person re-id in video (or multi-shot images in general) by uniquely exploiting person *tracklet labelling* obtained by an *unsupervised* tracklet formation (sampling) mechanism[2] *without* any ID labelling of the training data (either cross-view or within-view). The TAUDL trains a person re-id model in an end-to-end manner in order to benefit from the inherent overall model optimisation advantages from deep learning. In the following, we first present a data sampling mechanism for unsupervised within-camera tracklet labelling (Sect. 3.1) and then describe our model design for cross-camera tracklet association by joint unsupervised deep learning (Sect. 3.2).

3.1 Unsupervised Within-View Tracklet Labelling

Given a large quantity of video data from multiple disjoint cameras, we can readily deploy existing pedestrian detection and tracking models [26,41,56,61], to extract person tracklets. In general, the space-time trajectory of a person in a single-camera view from a public scene is likely to be fragmented into an arbitrary number of short tracklets due to imperfect tracking and background clutter. Given a large number of person tracklets per camera, we want to annotate them for deep re-id model learning in an *unsupervised* manner without any manual

[2] Although object tracklets can be generated by any independent single-camera-view multi-object tracking (MOT) models widely available today, a conventional MOT model is *not* end-to-end optimised for cross-camera tracklet association.

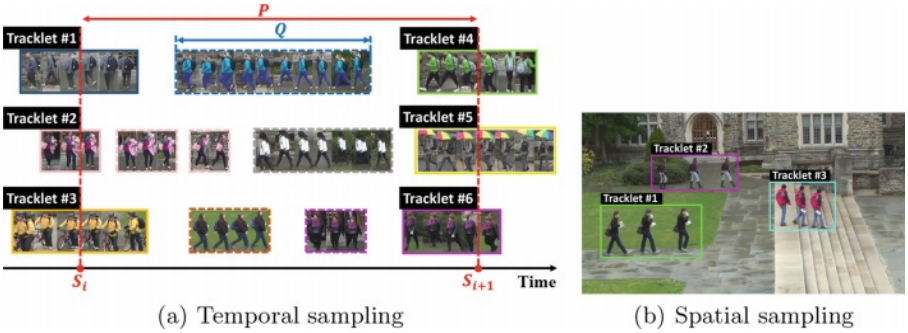

Fig. 2. An illustration of the Sparse Space-Time Tracklet sampling and annotating method for unsupervised tracklet labelling. Solid box: Sampled tracklets; Dashed box: Non-sampled tracklets; Each colour represents a distinct person ID. (a) Two time instances (S_i and S_{i+1} indicated by vertical lines) of temporal sampling are shown with a time gap P greater than the common transit time Q of a camera view. (b) Three spatially sparse tracklets are formed at a given temporal sampling instance.

identity verification on tracklets. To this end, we need an automatic tracklet labelling method to minimise the person ID duplication (i.e. multiple tracklet labels corresponding the same person ID label) rate among these labelled tracklets. To this end, we propose a **Sparse Space-Time Tracklet** (SSTT) sampling and label assignment method.

Our SSTT method is built on three observations typical in surveillance videos: **(1)** For most people, re-appearing in a camera view is rare during a short time period. As such, the dominant factor for causing person tracklet duplication (of the same ID) in auto-generated person tracklets is trajectory fragmentation, and if we assign every tracklet with a distinct label. To address this problem, we perform sparse temporal sampling of tracklets (Fig. 2(a)) as follows: (i) At the i-th temporal sampling instance corresponding to a time point S_i, we retrieve all tracklets at time S_i and annotate each tracklet with a distinct label. This is based on the factor that **(2)** people co-occurring at the same time in a single-view but at different spatial locations should have distinct ID labels. (ii) Given a time gap P, the next $((i+1)$-th) temporal sampling and label assignment is repeated, where P controls the sparsity of the temporal sampling rate. Based on observation **(3)** that most people in a public scene travel through a single camera view in a common time period $Q < P$, it is expected that at most one tracklet per person can be sampled at such a sparse temporal sampling rate (assuming no re-appearing once out of the same camera view). Consequently, we can significantly reduce the ID duplication even in highly crowded scenes with greater degrees of trajectory fragmentation.

To further mitigate the negative effect of inaccurate person detection and tracking at each temporal sampling instance, we further impose a sparse spatial sampling constraint – only selecting the co-occurring tracklets distantly distributed over the scene space (Fig. 2(b)). In doing so, the tracklet labels are more

likely to be of independent person identities with minimum ID duplications in each i-th temporal sampling instance.

By deploying this SSTT tracklet labelling method in each camera view, we can obtain an independent set of labelled tracklets $\{\boldsymbol{S}_i, y_i\}$ per-camera in a camera network, where each tracklet contains a varying number of person bounding boxes as $\boldsymbol{S} = \{\boldsymbol{I}_1, \boldsymbol{I}_2, \cdots\}$. Our objective is to use these SSTT labelled tracklets for optimising a cross-view person re-id deep learning model without any cross-view ID labelled pairwise training data.

3.2 Unsupervised Tracklet Association

Given per-camera independently-labelled tracklets $\{\boldsymbol{S}_i, y_i\}$ generated by SSTT, we perform *tracklet label re-id discriminative learning* without person ID labels in a conventional classification deep learning framework. To that end, we formulate a **Tracklet Association Unsupervised Deep Learning** (TAUDL) model. The overall design of our TAUDL architecture is shown in Fig. 1. The TAUDL contains two model components: **(I)** *Per-Camera Tracklet Discrimination Learning* with the aim to optimise "local" (within-camera) tracklet label discrimination for facilitating cross-camera tracklet association given independently created tracklet label spaces in different camera views. **(II)** *Cross-Camera Tracklet Association Learning* with the aim to maximise "global" (cross-camera) tracklet label association. The two components integrate as a whole in a single deep learning network architecture, learn jointly and mutually benefit each other in an incremental end-to-end manner.

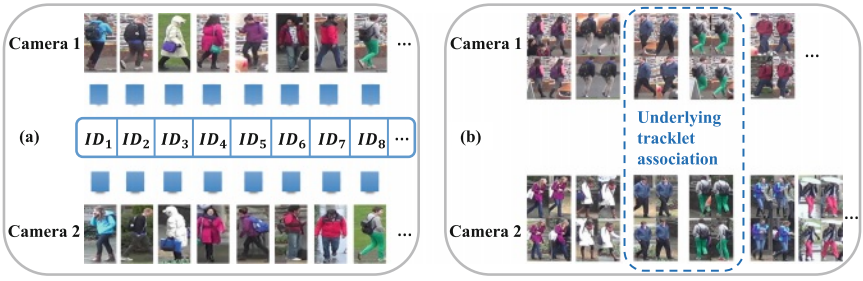

Fig. 3. Comparing (a) Fine-grained *explicit instance-level* cross-view ID labelled image pairs for supervised person re-id model learning and (b) Coarse-grained *latent group-level* cross-view tracklet (a multi-shot group) label correlation for ID label-free (unsupervised) person re-id learning using TAUDL.

(I) Per-Camera Tracklet Discrimination Learning. For accurate cross-camera tracklet association, it is important to formulate a robust image feature representation for describing the person appearance of each tracklet that helps cross-view person re-id association. However, it is sub-optimal to achieve

"local" per-camera tracklet discriminative learning using only per-camera independent tracklet labels without "global" cross-camera tracklet correlations. We wish to optimise jointly both local tracklet within-view discrimination and global tracklet cross-view association. To that end, we design a Per-Camera Tracklet Discrimination (PCTD) learning algorithm. Our key idea is that, instead of relying on the conventional fine-grained *explicit instance-level* cross-view ID pairwise supervised learning (Fig. 3(a)), we learn to maximise coarse-grained *latent group-level* cross-camera tracklet association by set correlation (Fig. 3(b)).

Specifically, we treat each individual camera view separately by optimising per-camera labelled tracklet discrimination as a classification task against the tracklet labels per-camera (not person ID labels). Therefore, we have a total of T different tracklet classification tasks each corresponding to a specific camera view. Importantly, we further formulate these T classification tasks in a multi-branch architecture design where every task shares the *same* feature representation whilst enjoys an individual classification branch (Fig. 1(b)). Conceptually, this model design is in a spirit of the multi-task learning principle [2,12].

Formally, given unsupervised training data $\{\boldsymbol{I}, y\}$ extracted from a camera view $t \in \{1, \cdots, T\}$, where $\boldsymbol{I}$ specifies a tracklet frame and $y \in \{1, \cdots, M_t\}$ the tracklet label (obtained as in Sect. 3.1) with a total of M_t different labels, we adopt the softmax Cross-Entropy (CE) loss function to optimise the corresponding classification task (the t-th branch). The CE loss on a training image sample $(\boldsymbol{I}, y)$ is computed as:

$$\mathcal{L}_{\text{ce}} = -\log\Big(\frac{\exp(\boldsymbol{W}_y^\top \boldsymbol{x})}{\sum_{k=1}^{M_t} \exp(\boldsymbol{W}_k^\top \boldsymbol{x})}\Big), \tag{1}$$

where $\boldsymbol{x}$ specifies the feature vector of $\boldsymbol{I}$ extracted by the task-shared feature representation component and $\boldsymbol{W}_y$ the y-th class prediction function parameters. Given a mini-batch, we compute the CE loss for each such training sample w.r.t. the respective tracklet label space and utilise their average to form the model learning supervision as:

$$\mathcal{L}_{\text{pctd}} = \frac{1}{N_{\text{bs}}} \sum_{t=1}^{T} \mathcal{L}_{\text{ce}}^t, \tag{2}$$

where $\mathcal{L}_{\text{ce}}^t$ denotes the CE loss summation of training samples from the t-th camera among a total of T and N_{bs} the batch size.

Discussion: In PCTD, the deep learning objective loss function (Eq. (1)) aims to optimise by supervised learning person tracklet discrimination *within* each camera view without any knowledge on *cross-camera* tracklet association. However, when jointly learning all the per-camera tracklet discrimination tasks together, the learned representation model is somewhat *implicitly* and *collectively* cross-view tracklet discriminative in a latent manner, due to the existence of cross-camera tracklet correlation. In other words, the shared feature representation is optimised *concurrently* to be discriminative for tracklet discrimination in multiple camera views, therefore propagating model discriminative learning

from per-camera to cross-camera. We will evaluate the effect of this model design in our experiments (Table 4).

(II) Cross-Camera Tracklet Association Learning. While the PCTD algorithm described above achieves somewhat global (all the camera views) tracklet discrimination implicitly, the learned model representation remains sub-optimal due to the lack of *explicitly* optimising cross-camera tracklet association at the fine-grained instance level. It is significantly harder to impose cross-view person re-id discriminative model learning without camera pairwise ID labels. To address this problem, we introduce a Cross-Camera Tracklet Association (CCTA) loss function. The CCTA loss is formulated based on the idea of *batch-wise incrementally aligning cross-view per tracklet feature distribution* in the shared multi-task learning feature space. Critically, CCTA integrates seamlessly with PCTD to jointly optimise model learning on discovering cross-camera tracklet association for person re-id in a single end-to-end batch-wise learning process.

Formally, given a mini-batch including a subset of tracklets $\{(\boldsymbol{S}_i^t, y_i^t)\}$ where $\boldsymbol{S}_i^t$ specifies the i-th tracklet from t-th camera view with the label y_i^t where tracklets in a mini-batch come from all the camera views, we want to establish for each in-batch tracklet a discriminative association with other tracklets from different camera views. In absence of person identity pairwise labelling as a learning constraint, we propose to align *similar* and *dissimilar* tracklets in each mini-batch given the up-to-date shared multi-task (multi-camera) feature representation from optimising PCTD. More specifically, for each tracklet $\boldsymbol{S}_i^t$, we first retrieve K cross-view nearest tracklets $\mathcal{N}_i^t$ in the feature space, with the remaining $\tilde{\mathcal{N}}_i^t$ considered as dissimilar ones. We then impose a soft discriminative structure constraint by encouraging the model to pull $\mathcal{N}_i^t$ close to $\boldsymbol{S}_i^t$ whilst to push away $\tilde{\mathcal{N}}_i^t$ from $\boldsymbol{S}_i^t$. Conceptually, this is a per-tracklet cross-view data structure distribution alignment. To achieve this, we formulate a CCTA deep learning objective loss for each tracklet $\boldsymbol{S}_i^t$ in a training mini-batch as:

$$\mathcal{L}_{\text{ccta}} = -\log \frac{\sum_{\boldsymbol{z}_k \in \mathcal{N}_i^t} \exp(-\frac{1}{2\sigma^2} \parallel \boldsymbol{s}_i^t - \boldsymbol{z}_k \parallel_2)}{\sum_{t'=1}^{T} \sum_{j=1}^{n_j} \exp(-\frac{1}{2\sigma^2} \parallel \boldsymbol{s}_i^t - \boldsymbol{s}_j^{t'} \parallel_2)}, \tag{3}$$

where n_j denotes the number of in-batch tracklets from j-th camera view, T the camera view number, σ a scaling parameter, $\boldsymbol{s}_i^t$ the up-to-date feature representation of the tracklet $\boldsymbol{S}_i^t$. Given the incremental iterative deep learning nature, we represent a tracklet $\boldsymbol{S}$ by the average of its in-batch frames' feature vectors on-the-fly. Hence, the tracklet representation is kept up-to-date without the need for maintaining external per-tracklet feature representations.

Discussion: The proposed CCTA loss formulation is conceptually similar to the Histogram Loss [44] in terms of distribution alignment. However, the Histogram Loss is a *supervised* loss that requires supervised label training data, whilst the CCTA is purely *unsupervised* and derived directly from feature similarity measures. CCTA is also related to the surrogate (artificially built) class based unsupervised deep learning loss formulations [4,5], by not requiring groundtruth class-labelled data in model training. Unlike CCTA without the need for creating

surrogate classes, the surrogate based models not only require additional global data clustering, but also are sensitive to the clustering quality and initial feature selection. Moreover, they do not consider the label distribution alignment across cameras and label spaces for which the CCTA loss is designed.

Joint Loss Function. After merging the CCTA and PCTD learning constraints, we obtain the final model objective function as:

$$\mathcal{L}_{\text{taudl}} = (1 - \lambda)\mathcal{L}_{\text{pctd}} + \lambda\mathcal{L}_{\text{ccta}}, \tag{4}$$

where λ is a weighting parameter estimated by cross-validation. Note that $\mathcal{L}_{\text{pctd}}$ is an average loss term at the tracklet individual image level whilst $\mathcal{L}_{\text{ccta}}$ at the tracklet group (set) level, both derived from the same training batch concurrently. As such, the overall TAUDL method naturally enables end-to-end deep model learning using the Stochastic Gradient Descent optimisation algorithm.

4 Experiments

Datasets. To evaluate the proposed TAUDL model, we tested both video (MARS [59], iLIDS-Video [50], PRID2011 [19]) and image (CUHK03 [29], Market-1501 [60], DukeMTMC [41,61]) based person re-id benchmarking datasets. In previous studies, these datasets were mostly evaluated separately. We consider since recent large sized image based re-id datasets were typically constructed by sampling person bounding boxes from video, these image datasets share similar characteristics of those video based datasets. We adopted the standard person re-id setting on training/test ID split and the test protocols (Table 1).

Fig. 4. Example cross-view matched image/tracklet pairs from (a) CUHK03, (b) Market-1501, (c) DukeMTMC, (d) PRID2011, (e) iLIDS-VID, (f) MARS.

Tracklet Label Assignment. For all six datasets, we cannot perform real SSTT tracklet sampling and label assignment due to no information available on spatial and temporal location w.r.t. the original video data. In our experiment, we instead conducted simulated SSTT to obtain the per-camera tracklet/image

Table 1. Dataset statistics and evaluation setting.

Dataset	# ID	# Train	# Test	# Images	# Tracklet
iLIDS-VID [50]	300	150	150	43,800	600
PRID2011 [19]	178	89	89	38,466	354
MARS [59]	1,261	625	636	1,191,003	20,478
CUHK03 [29]	1,467	767	700	14,097	0
Market-1501 [60]	1,501	751	750	32,668	0
DukeMTMC [41]	1,812	702	1,110	36,411	0

labels. For all datasets, we assume no re-appearing subjects per camera (very rare in these datasets) and sparse spatial sampling. As both iLIDS-VID and PRID2011 provide only one tracklet per ID per camera (i.e. no fragmentation), it is impossible to have per-camera ID duplication. Therefore, each tracklet is assigned a unique label. The MARS gives multiple tracklets per ID per camera. Based on SSTT, at most only one tracklet can be sampled for each ID per camera (see Sect. 3.1). Therefore, a MARS tracklet per ID per camera was randomly selected and assigned a label. For all image based datasets, we assume all images per ID per camera were drawn from a single tracklet, same as in iLIDS-VID and PRID2011. The same tracklet label assignment procedure was adopted as above.

Performance Metrics. We use the common cumulative matching characteristic (CMC) and mean Average Precision (mAP) metrics [60].

Implementation Details. We adopted an ImageNet pre-trained ResNet-50 [17] as the backbone in evaluating the proposed TAUDL method. We set the feature dimension of the camera-shared representation space derived on top of ResNet-50 to 2,048. Each camera-specific branch contains one FC classification layer. Person images are resized to 256 $\times$ 128 for all datasets. To ensure that each batch has the capacity of containing person images from all cameras, we set the batch size to 384 for all datasets. For balancing the model learning speed over different cameras, we randomly selected the same number of training frame images per camera when sampling each mini-batch. We adopted the Adam optimiser [23] with the initial learning rate of 3.5×10^{-4}. We empirically set $\lambda = 0.7$ for Eq. (4), $\sigma = 2$ for Eq. (3), and $K = T/2$ (T is the number of cameras) for cross-view nearest tracklets $\mathcal{N}_i^t$ in Eq. (3) for all the experiments.

4.1 Comparisons to State-of-the-Arts

We compared two different sets of state-of-the-art methods on image and video re-id datasets, due to the independent studies on them in the literature.

Unsupervised Person Re-ID on Image Datasets. Table 2 shows the unsupervised re-id performance of the proposed TAUDL and 10 state-of-the-art methods including 3 hand-crafted feature based methods (Dic [25], ISR [32], RKSL [48])

Table 2. Unsupervised re-id on image datasets. $1^{st}/2^{nd}$ best results are in **red**/**blue**.

Dataset	CUHK03 [29]		Market-1501 [60]		DukeMTMC [61]	
Metric (%)	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
Dic [25]	36.5	-	50.2	22.7	-	-
ISR [32]	38.5	-	40.3	14.3	-	-
RKSL [48]	34.8	-	34.0	11.0	-	-
SAE [27]	30.5	-	42.4	16.2	-	-
JSTL [52]	33.2	-	44.7	18.4	-	-
AML [53]	31.4	-	44.7	18.4	-	-
UsNCA [40]	29.6	-	45.2	18.9	-	-
CAMEL [55]	**39.4**	-	54.5	26.3	-	-
PUL [13]	-	-	44.7	20.1	30.4	16.4
TJ-AIDL [49]	-	-	**58.2**	**26.5**	**44.3**	**23.0**
TAUDL	**44.7**	**31.2**	**63.7**	**41.2**	**61.7**	**43.5**
GCS [6](*Supervised*)	88.8	97.2	93.5	81.6	84.9	69.5

and 7 auxiliary knowledge (identity/attribute) transfer based models (AE [27], AML [53], UsNCA [40], CAMEL [55], JSTL [52], PUL [13], TJ-AIDL [49]). These results show: **(1)** Among existing methods, the knowledge transfer based method is superior, e.g. on CUHK03, Rank-1 39.4% by CAMEL vs. 36.5% by Dic; On Market-1501, 58.2% by TJ-AIDL vs. 50.2% by Dic. To that end, CAMEL benefits from learning on 7 different person re-id datasets of diverse domains (CUHK03 [29], CUHK01 [28], PRID [19], VIPeR [16], 3DPeS [3], i-LIDS [39], Shinpuhkan [21]) including a total of 44,685 images and 3,791 identities; TJ-AIDL utilises labelled Market-1501 (750 IDs and 27 attribute classes) or DukeMTMC (702 IDs and 23 attribute classes) as source training data. **(2)** Our new model TAUDL outperforms all competitors with significant margins. For example, the Rank-1 margin by TAUDL over TJ-AIDL is 5.5% (63.7–58.2) on Market-1501 and 17.4% (61.7–44.3) on DukeMTMC. Moreover, it is worth pointing out that TAUDL dose not benefit from any additional labelled source domain training data as compared to TJ-AIDL. TAUDL is potentially more scalable due to no need to consider source and target domains similarities. **(3)** Our TAUDL is simpler to train with a simple end-to-end model learning, as compared to the alternated deep CNN training and clustering required by PUL and a two-stage model training of TJ-AIDL. These results show both the performance advantage and model design superiority of the proposed TAUDL model over a wide variety of state-of-the-art re-id models.

Unsupervised Person Re-ID on Video Datasets. We compared the proposed TAUDL with six state-of-the-art unsupervised video person re-id models. Unlike TAUDL, all these existing models are not end-to-end deep learning methods with either hand-crafted or separately trained deep features as model input. Table 3 shows that TAUDL outperforms all existing video-based person re-id

Table 3. Unsupervised re-id on video datasets. $1^{st}/2^{nd}$ best results are in **red**/**blue**.

Dataset	PRID2011 [19]			iLIDS-VID [50]			MARS [59]			
Metric (%)	R1	R5	R20	R1	R5	R20	R1	R5	R20	mAP
DTW [37]	41.7	67.1	90.1	31.5	62.1	82.4	-	-	-	-
GRDL [24]	41.6	76.4	89.9	25.7	49.9	77.6	19.3	33.2	46.5	9.56
UnKISS [22]	58.1	81.9	96.0	35.9	**63.3**	**83.4**	22.3	37.4	53.6	10.6
SMP [35]	**80.9**	**95.6**	**99.4**	**41.7**	**66.3**	80.7	23.9	35.8	44.9	10.5
DGM+MLAPG [54]	**73.1**	**92.5**	**99.0**	**37.1**	61.3	82.0	24.6	42.6	57.2	11.8
DGM+IDE [54]	56.4	81.3	96.4	36.2	62.8	**82.7**	**36.8**	**54.0**	**68.5**	**21.3**
TAUDL	49.4	78.7	98.9	26.7	51.3	82.0	**43.8**	**59.9**	**72.8**	**29.1**
QAN [34](*Supervised*)	90.3	98.2	100.0	68.0	86.8	97.4	73.7	84.9	91.6	51.7

models on the large scale video dataset MARS, e.g. by a Rank-1 margin of 7.0% (43.8–36.8) over the best competitor DGM+IDE (which additionally using the ID label information of one camera view for model initialisation). However, TAUDL is inferior than some of the existing models on the two small benchmarks iLIDS-VID (300 training tracklets) and PRID2011 (178 training tracklets), in comparison to its performance on the MARS benchmark (8,298 training tracklets). This shows that TAUDL does need sufficient tracklet data from larger video datasets in order to have its performance advantage. As the tracklet data required are not manually labelled, this requirement is not a hindrance to its scalability to large scale data. Quite the contrary, TAUDL works the best when large scale unlabelled video data is available. A model would benefit particularly from pre-training using TAUDL on large auxiliary unlabelled video data from similar camera viewing conditions.

4.2 Component Analysis and Discussions

Effectiveness of Per-Camera Tracklet Discrimination. The PCTD component was evaluated by comparing a baseline that treats all cameras together by concatenating per-camera tracklet label sets and deploying the Cross-Entropy loss to learn a unified classification task. We call this baseline Joint-Camera Classification (JCC). In this analysis, we do not consider the cross-camera tracklet association component for a clear evaluation. Table 4 shows that our PCTD design is significantly superior over the JCC learning algorithm, e.g. achieving Rank-1 gain of 4.0%, 34.6%, 36.3%, and 19.9% on CUHK03, Market-1501, DukeMTMC, and MARS respectively. This verifies the modelling advantages of the proposed per-camera tracklet discrimination learning scheme on the unsupervised tracklet labels in inducing cross-view re-id discriminative feature learning.

Effectiveness of Cross-Camera Tracklet Association. The CCTA learning component was evaluated by testing the performance drop after eliminating it. Table 5 shows a significant performance benefit from this model component,

e.g. a Rank-1 boost of 10.9%, 11.6%, 10.5%, and 5.8% on CUHK03, Market-1501, DukeMTMC, and MARS respectively. This validates the importance of modelling the correlation across cameras in discriminative optimisation and the effectiveness of our CCTA deep learning objective loss formulation in an end-to-end manner. Additionally, this also suggests the effectiveness of the PCTD model component in facilitating the cross-view identity discrimination learning by providing re-id sensitive features in a joint incremental learning manner.

Table 4. Effect of Per-Camera Tracklet Discrimination (PCTD) learning.

Dataset	CUHK03 [29]		Market-1501 [60]		DukeMTMC [41]		MARS [59]	
Metric(%)	R1	mAP	R1	mAP	R1	mAP	R1	mAP
JCC	29.8	12.5	17.5	7.9	14.9	3.5	18.1	13.1
PCTD	**33.8**	**18.9**	**52.1**	**26.6**	**51.2**	**32.9**	**38.0**	**23.9**

Table 5. Effect of Cross-Camera Tracklet Association (CCTA)

Dataset	CUHK03 [29]		Market-1501 [60]		DukeMTMC [61]		MARS [59]	
CCTA	R1	mAP	R1	mAP	R1	mAP	R1	mAP
✗	33.8	18.9	52.1	26.6	51.2	32.9	38.0	23.9
✓	**44.7**	**31.2**	**63.7**	**41.2**	**61.7**	**43.5**	**43.8**	**29.1**

Model Robustness Analysis. Finally, we performed an analysis on model robustness against person ID duplication rates in tracklet labelling. We conducted a controlled evaluation on MARS where multiple tracklets per ID per camera are available for setting simulation. Recall that the ID duplication may mainly come with imperfect temporal sampling due to trajectory fragmentation and when some people stay in the same camera view for a longer time period than the temporal sampling gap. To simulate such a situation, we assume a varying percentage (10%∼50%) of IDs per camera have two random tracklets sampled and annotated with different tracklet labels. More tracklets per ID per camera are likely to be sampled, which can make this analysis more complex due to the interference from the number of duplicated person IDs. Table 6 shows that our TAUDL model is robust against the ID duplication rate, e.g. with only a Rank-1 drop of 3.1% given as high as 50% per-camera ID duplication rate. In reality, it is not too hard to minimise ID duplication rate among tracklets (Sect. 3.1), e.g. conducting very sparse sampling over time and space. Note, we do not care about exhaustive sampling of all the tracklets from video in a given time period. The model learning benefits from very sparse and diverse tracklet sampling from a large pool of unlabelled video data.

The robustness of our TAUDL comes with two model components: **(1)** The model learning optimisation is not only subject to a single per-camera tracklet label constraint, but also concurrently to the constraints of all cameras. This facilitates optimising cross-camera tracklet association globally across all cameras in a common space, due to the Per-Camera Tracklet Discrimination learning mechanism (Eq. (2)). This provides model learning tolerance against per-camera tracklet label duplication errors. **(2)** The cross-camera tracklet association learning is designed as a feature similarity based "soft" objective learning constraint (Eq. (3)), without a direct dependence on the tracklet ID labels. Therefore, the ID duplication rate has little effect on this objective loss constraint.

Table 6. Model robustness analysis on varying ID duplication rates on MARS [59].

ID duplication rate (%)	Rank-1	Rank-5	Rank-10	Rank-20	mAP
0	**43.8**	**59.9**	**66.0**	**72.8**	**29.1**
10	42.8	59.7	65.5	71.6	28.3
20	42.2	58.8	64.7	70.6	27.4
30	41.6	57.9	64.5	69.7	26.7
50	40.7	57.0	63.4	69.6	25.6

5 Conclusions

In this work, we presented a novel *Tracklet Association Unsupervised Deep Learning* (TAUDL) model for unsupervised person re-identification using unsupervised person tracklet data extracted from videos, therefore eliminating the tedious and exhaustive manual labelling required by all supervised learning based re-id model learning. This enables TAUDL to be much more scalable to real-world re-id deployment at large scale video data. In contrast to most existing re-id methods that either require exhaustively pairwise labelled training data for every camera pair or assume the availability of additional labelled source domain training data for target domain adaptation, the proposed TAUDL model is capable of end-to-end deep learning a discriminative person re-id model from scratch on totally unlabelled tracklet data. This is achieved by optimising jointly both the Per-Camera Tracklet Discrimination loss function and the Cross-Camera Tracklet Association loss function in a single end-to-end deep learning framework. To our knowledge, this is the first completely unsupervised learning based re-id model without any identity labels for model learning, neither pairwise cross-view image pair labelling nor single-view image identity class labelling. Extensive comparative evaluations were conducted on six image and video based re-id benchmarks to validate the advantages of the proposed TAUDL model over a wide range of state-of-the-art unsupervised and domain adaptation re-id methods. We also conducted in-depth TAUDL model component evaluation and robustness test to give insights on model performance advantage and model learning stability.

Acknowledgments. This work is partially supported by the China Scholarship Council, Vision Semantics Limited, National Natural Science Foundation of China (Project No. 61401212), the Key Technology Research and Development Program of Jiangsu Province (Project No. BE2015162), the Science and Technology Support Project of Jiangsu Province (Project No. BE2014714), Royal Society Newton Advanced Fellowship Programme (NA150459), and Innovate UK Industrial Challenge Project on Developing and Commercialising Intelligent Video Analytics Solutions for Public Safety (98111-571149).

References

1. Ahmed, E., Jones, M., Marks, T.K.: An improved deep learning architecture for person re-identification. In: CVPR (2015)
2. Ando, R.K., Zhang, T.: A framework for learning predictive structures from multiple tasks and unlabeled data. JMLR **6**, 1817–1853 (2005)
3. Baltieri, D., Vezzani, R., Cucchiara, R.: 3DPeS: 3D people dataset for surveillance and forensics. In: J-HGBU (2011)
4. Bautista, M.A., Sanakoyeu, A., Ommer, B.: Deep unsupervised similarity learning using partially ordered sets. In: CVPR (2017)
5. Bautista, M.A., Sanakoyeu, A., Tikhoncheva, E., Ommer, B.: CliqueCNN: deep unsupervised exemplar learning. In: NIPS (2016)
6. Chen, D., Xu, D., Li, H., Sebe, N., Wang, X.: Group consistent similarity learning via deep CRF for person re-identification. In: CVPR (2018)
7. Chen, W., Chen, X., Zhang, J., Huang, K.: Beyond triplet loss: a deep quadruplet network for person re-identification. In: CVPR (2017)
8. Chen, Y., Zhu, X., Gong, S.: Person re-identification by deep learning multi-scale representations. In: ICCV Workshop (2017)
9. Chen, Y.C., Zhu, X., Zheng, W.S., Lai, J.H.: Person re-identification by camera correlation aware feature augmentation. IEEE TPAMI **40**(2), 392–408 (2018)
10. Cheng, D., Gong, Y., Zhou, S., Wang, J., Zheng, N.: Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In: CVPR (2016)
11. Cho, Y.J., Yoon, K.J.: Improving person re-identification via pose-aware multi-shot matching. In: CVPR (2016)
12. Evgeniou, T., Pontil, M.: Regularized multi-task learning. In: SIGKDD (2004)
13. Fan, H., Zheng, L., Yang, Y.: Unsupervised person re-identification: clustering and fine-tuning. arXiv preprint arXiv:1705.10444 (2017)
14. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person re-identification by symmetry-driven accumulation of local features. In: CVPR (2010)
15. Gong, S., Cristani, M., Yan, S., Loy, C.C.: Person Re-identification. Springer, London (2014). https://doi.org/10.1007/978-1-4471-6296-4
16. Gray, D., Tao, H.: Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008. LNCS, vol. 5302, pp. 262–275. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-88682-2_21
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
18. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 (2017)

19. Hirzer, M., Beleznai, C., Roth, P.M., Bischof, H.: Person re-identification by descriptive and discriminative classification. In: Heyden, A., Kahl, F. (eds.) SCIA 2011. LNCS, vol. 6688, pp. 91–102. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-21227-7_9
20. Jiao, J., Zheng, W.S., Wu, A., Zhu, X., Gong, S.: Deep low-resolution person re-identification. In: AAAI (2018)
21. Kawanishi, Y., Wu, Y., Mukunoki, M., Minoh, M.: Shinpuhkan 2014: A multi-camera pedestrian dataset for tracking people across multiple cameras. In: FCV (2014)
22. Khan, F.M., Bremond, F.: Unsupervised data association for metric learning in the context of multi-shot person re-identification. In: AVSS (2016)
23. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
24. Kodirov, E., Xiang, T., Fu, Z., Gong, S.: Person re-identification by unsupervised l_1 graph learning. In: ECCV (2016)
25. Kodirov, E., Xiang, T., Gong, S.: Dictionary learning with iterative Laplacian regularisation for unsupervised person re-identification. In: BMVC (2015)
26. Leal-Taixé, L., Milan, A., Reid, I., Roth, S., Schindler, K.: Motchallenge 2015: towards a benchmark for multi-target tracking. arXiv preprint arXiv:1504.01942 (2015)
27. Lee, H., Ekanadham, C., Ng, A.Y.: Sparse deep belief net model for visual area v2. In: NIPS (2008)
28. Li, W., Zhao, R., Wang, X.: Human reidentification with transferred metric learning. In: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) ACCV 2012. LNCS, vol. 7724, pp. 31–44. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-37331-2_3
29. Li, W., Zhao, R., Xiao, T., Wang, X.: DeepReID: deep filter pairing neural network for person re-identification. In: CVPR (2014)
30. Li, W., Zhu, X., Gong, S.: Person re-identification by deep joint learning of multi-loss classification. In: IJCAI (2017)
31. Li, W., Zhu, X., Gong, S.: Harmonious attention network for person re-identification. In: CVPR (2018)
32. Lisanti, G., Masi, I., Bagdanov, A.D., Del Bimbo, A.: Person re-identification by iterative re-weighted sparse ranking. IEEE TPAMI **37**(8), 1629–1642 (2015)
33. Liu, X., Song, M., Tao, D., Zhou, X., Chen, C., Bu, J.: Semi-supervised coupled dictionary learning for person re-identification. In: CVPR (2014)
34. Liu, Y., Yan, J., Ouyang, W.: Quality aware network for set to set recognition. In: CVPR (2017)
35. Liu, Z., Wang, D., Lu, H.: Stepwise metric promotion for unsupervised video person re-identification. In: ICCV (2017)
36. Loy, C., Xiang, T., Gong, S.: Time-delayed correlation analysis for multi-camera activity understanding. IJCV **90**(1), 106–129 (2010)
37. Ma, X., Zhu, X., Gong, S., Xie, X., Hu, J., Lam, K.M., Zhong, Y.: Person re-identification by unsupervised video matching. Pattern Recogn. **65**, 197–210 (2017)
38. Peng, P., Xiang, T., Wang, Y., Pontil, M., Gong, S., Huang, T., Tian, Y.: Unsupervised cross-dataset transfer learning for person re-identification. In: CVPR (2016)
39. Prosser, B.J., Zheng, W.S., Gong, S., Xiang, T.: Person re-identification by support vector ranking. In: BMVC (2010)
40. Qin, C., Song, S., Huang, G., Zhu, L.: Unsupervised neighborhood component analysis for clustering. Neurocomputing **168**, 609–617 (2015)

41. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: ECCV Workshop (2016)
42. Su, C., Zhang, S., Xing, J., Gao, W., Tian, Q.: Deep attributes driven multi-camera person re-identification. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 475–491. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_30
43. Subramaniam, A., Chatterjee, M., Mittal, A: Deep neural networks with inexact matching for person re-identification. In: NIPS (2016)
44. Ustinova, E., Lempitsky, V.: Learning deep embeddings with histogram loss. In: NIPS (2016)
45. Wang, F., Zuo, W., Lin, L., Zhang, D., Zhang, L.: Joint learning of single-image and cross-image representations for person re-identification. In: CVPR (2016)
46. Wang, H., Gong, S., Xiang, T.: Unsupervised learning of generative topic saliency for person re-identification. In: BMVC (2014)
47. Wang, H., Zhu, X., Gong, S., Xiang, T.: Person re-identification in identity regression space. IJCV (2018)
48. Wang, H., Zhu, X., Xiang, T., Gong, S.: Towards unsupervised open-set person re-identification. In: ICIP (2016)
49. Wang, J., Zhu, X., Gong, S., Li, W.: Transferable joint attribute-identity deep learning for unsupervised person re-identification. In: CVPR (2018)
50. Wang, T., Gong, S., Zhu, X., Wang, S.: Person re-identification by video ranking. In: ECCV (2014)
51. Wang, T., Gong, S., Zhu, X., Wang, S.: Person re-identification by discriminative selection in video ranking. IEEE TPAMI **38**(12), 2501–2514 (2016)
52. Xiao, T., Li, H., Ouyang, W., Wang, X.: Learning deep feature representations with domain guided dropout for person re-identification. In: CVPR (2016)
53. Ye, J., Zhao, Z., Liu, H.: Adaptive distance metric learning for clustering. In: CVPR (2007)
54. Ye, M., Ma, A.J., Zheng, L., Li, J., Yuen, P.C.: Dynamic label graph matching for unsupervised video re-identification. In: ICCV (2017)
55. Yu, H.X., Wu, A., Zheng, W.S.: Cross-view asymmetric metric learning for unsupervised person re-identification. In: ICCV (2017)
56. Zhang, S., Benenson, R., Omran, M., Hosang, J., Schiele, B.: How far are we from solving pedestrian detection? In: CVPR (2016)
57. Zhang, Y., Xiang, T., Hospedales, T.M., Lu, H.: Deep mutual learning. In: CVPR (2018)
58. Zhao, R., Oyang, W., Wang, X.: Person re-identification by saliency learning. IEEE TPAMI **39**(2), 356–370 (2017)
59. Zheng, L., et al.: Mars: a video benchmark for large-scale person re-identification. In: ECCV (2016)
60. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: a benchmark. In: CVPR (2015)
61. Zheng, Z., Zheng, L., Yang, Y.: Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. In: ICCV (2017)
62. Zhu, X., Wu, B., Huang, D., Zheng, W.S.: Fast openworld person re-identification. In: IEEE TIP, pp. 2286–2300 (2017)