Deep Association Learning for Unsupervised Video Person Re-identification

Yanbei Chen[1]
yanbei.chen@qmul.ac.uk
Xiatian Zhu[2]
eddy@visionsemantics.com
Shaogang Gong[1]
s.gong@qmul.ac.uk

[1] Computer Vision Group,
School of Electronic Engineering and
Computer Science,
Queen Mary University of London,
London E1 4NS, UK

[2] Vision Semantics Ltd.,
London E1 4NS, UK.

Abstract

Deep learning methods have started to dominate the research progress of video-based person re-identification (re-id). However, existing methods mostly consider supervised learning, which requires exhaustive manual efforts for labelling cross-view pairwise data. Therefore, they severely lack scalability and practicality in real-world video surveillance applications. In this work, to address the video person re-id task, we formulate a novel *Deep Association Learning* (DAL) scheme, the first end-to-end deep learning method using none of the identity labels in model initialisation and training. DAL learns a deep re-id matching model by jointly optimising two margin-based association losses in an end-to-end manner, which effectively constrains the association of each frame to the best-matched intra-camera representation and cross-camera representation. Existing standard CNNs can be readily employed within our DAL scheme. Experiment results demonstrate that our proposed DAL significantly outperforms current state-of-the-art unsupervised video person re-id methods on three benchmarks: PRID 2011, iLIDS-VID and MARS.

1 Introduction

Person re-identification (re-id) aims to match persons across disjoint camera views distributed at different locations [13]. While most recent re-id methods rely on static images [2, 6, 21, 22, 23, 32, 34, 35, 39, 47, 48, 52], video-based re-id has gained increasing attention [16, 28, 37, 38, 40, 41, 44, 45, 45, 50, 51] due to the rich space-time information inherently carried in the video tracklets. A video tracklet is a sequence of images that captures rich variations of the same person in terms of occlusion, background cluster, viewpoint, human poses, etc, which can naturally be used as informative data sources for person re-id. The majority of current techniques in video person re-id consider the supervised learning context, which imposes a strong assumption on the availability of identity (ID) labels for every camera pair therefore allowing more powerful and discriminative re-id models to be learned when given relatively small-sized training data. However, supervised learning methods are weak in scaling to real-world deployment beyond the labelled training data domains. In practice, exhaustive manual annotation at every camera pair is not only prohibitively expensive

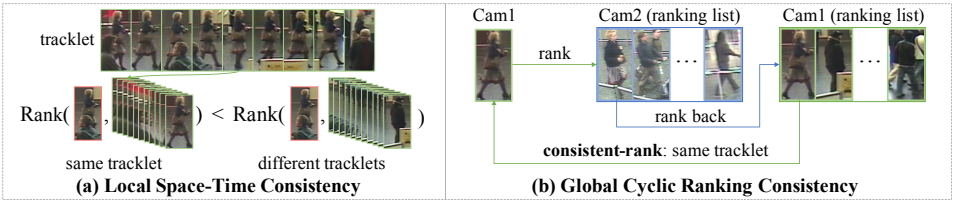


Figure 1: Two types of consistency in our Deep Association Learning scheme. (a) Local space-time consistency: Most images from the same tracklet generally depict the same person. (b) Global cyclic ranking consistency: Two tracklets from different cameras are highly associated if they are *mutually* the nearest neighbour returned by a cross-view ranking.

for a large identity population across a large camera network, but it is also implausible due to insufficient designated persons reappearing in every camera pair. In this regard, unsupervised video re-id is a more realistic task that is worth studying to improve the scalability of re-id models in practical use.

Unsupervised learning methods [18, 25, 26, 27, 36, 42] are particularly essential when the re-id task needs to be performed on a large amount of unlabelled video surveillance data cumulated continuously over time, whilst the pairwise ID labels cannot be easily acquired for supervised model learning. Due to the inherent nature of unsupervised learning, existing methods suffer from significant performance degradations when compared to supervised learning methods in video person re-id. For instance, the state-of-the-art rank-1 re-id matching rate on MARS [45] is only 36.8% by unsupervised learning [42], as compared to 82.3% by supervised learning [20]. In fact, even the latest video-based unsupervised learning models [26, 42] for person re-id still lack a principled mechanism to explore the more powerful representation-learning capabilities of deep Convolutional Neural Networks (CNNs) [3] for jointly learning an expressive embedding representation and a discriminative re-id matching model in an end-to-end manner. It is indeed not straightforward to formulate a deep learning scheme for unsupervised video-based person re-id due to: (1) The general supervised learning nature of deep CNN networks: most deep learning objectives are formulated on labelled training data; (2) The cross-camera variations of the same-ID tracklet pairs from disjoint camera views and the likelihood of different people being visually similar in public space, which collectively render the nearest-neighbour distance measure unreliable to capture the cross-view person identity matching for guiding the model learning.

In this work, we aim to tackle the task of unsupervised video person re-id by an end-to-end optimised deep learning scheme without utilising any ID labels. Towards this aim, we formulate a novel *unsupervised* **Deep Association Learning** (DAL) scheme designed specifically to explore two types of *consistency*, including (1) *local space-time consistency* within each tracklet from the same camera view, and (2) *global cyclic ranking consistency* between tracklets across disjoint camera views (Figure 1). In particular, we define two margin-based association losses, with one derived from the intra-camera tracklet representation updated incrementally on account of the *local space-time consistency*, and the other derived from the cross-camera representation learned continuously based on the *global cyclic ranking consistency*. Importantly, this scheme enables the deep model to start with learning from the local consistency, whilst incrementally self-discovering more cross-camera highly associated tracklets subject to the global consistency for progressively enhancing discriminative feature learning. Overall, our DAL scheme imposes batch-wise self-supervised learning cycles to eliminate the need for manual labelled supervision in the course of model training.

Our contribution is three-fold: **(I)** We propose for the first time an end-to-end deep learning scheme for unsupervised video person re-id without imposing any human knowledge on identity information. **(II)** We formulate a novel *Deep Association Learning* (DAL) scheme, with two discriminative association losses derived from (1) *local space-time consistency* within each tracklet and (2) *global cyclic ranking consistency* between tracklets across disjoint camera views. Our DAL loss formulation allows typical deep CNNs to be readily trained by standard stochastic gradient descent algorithms. **(III)** Extensive experiments demonstrate the advantages of DAL over the state-of-the-art unsupervised video person re-id methods on three benchmark datasets: PRID2011 [16], iLIDS-VID [37], and MARS [45].

2 Related Work

Unsupervised Video-based Person Re-identification has started to attract increasing research interest recently [18, 19, 26, 27, 42]. The commonality of most existing methods is to discover the matching correlations between tracklets across cameras. For example, Ma et al. [27] formulate a time shift dynamic warping model to automatically pair cross-camera tracklets by matching partial segments of each tracklet generated over all time shifts. Ye et al. [42] propose a dynamic graph matching method to mine the cross-camera labels for iteratively learning a discriminative distance metric model. Liu et al. [26] develop a stepwise metric learning method to progressively estimate the cross-camera labels; but it requires stringent video filtering to obtain one tracklet per ID per camera for discriminative model initialisation. The proposed Deep Association Learning (DAL) method in this work differs significantly from previous works in three aspects: (1) Unlike [26, 27], our DAL does not require additional manual effort to select tracklets for model initialisation, which results in better scalability to large-scale video data. (2) All existing methods rely on a good external feature extractor for metric learning; while our DAL jointly learns a re-id matching model with discriminative representation in a fully end-to-end manner. (3) Our DAL uniquely utilises the intra-camera local space-time consistency and cross-camera global cyclic ranking consistency to formulate the learning objective with a relatively low computational cost.

Deep Metric Learning aims to learn a nonlinear mapping that transforms input images into a feature representation space, in which the distances within the same class are enforced to be small whilst the distances between different classes are maintained large. A variety of deep distance metric learning methods have been proposed to solve the person re-id problem [2, 4, 5, 7, 10, 15, 21, 24, 28, 34, 40, 43], among which the most popular learning constraint is pairwise comparison [21, 43] or triplet comparison [15, 29, 30] (also known as relative distance comparison [10, 46]). For pairwise comparison, a binary classification learning objective [2, 21] or a Siamese network with a similarity measure objective [28, 40, 43] is typically adopted to learn a nonlinear mapping that outputs pairwise similarity scores. For triplet comparison, a margin-based hinge loss with a batch construction strategy for triplet generation [10, 15] is often deployed to maximise the relative distance between matched pairs and unmatched pairs of inputs. As opposed to most supervised deep metric learning methods in person re-id, our DAL learns a deep embedding representation in an unsupervised fashion. Instead of grounding the learning objective based on pairwise or triple-wise comparison between a few labelled samples, e.g., three samples as a triplet, our DAL uniquely learns two set of anchors as the intra-camera and cross-camera tracklet representations, which allows to measure the pairwise similarities between each image frame and all the tracklet representations to formulate the unsupervised learning objectives.

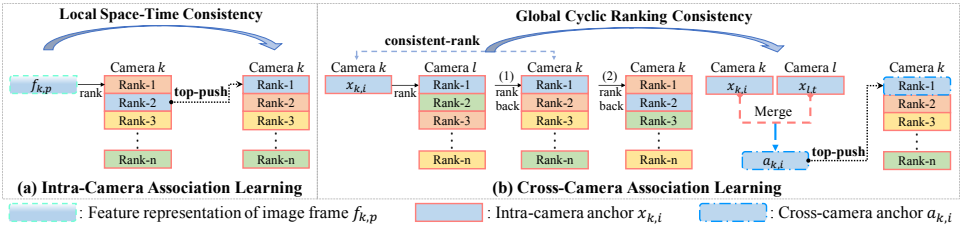


Figure 2: Illustration of Deep Association Learning: (a) Intra-camera association learning based on the local space-time consistency within tracklets (Sec. 3.1). (2) Cross-camera association learning based on the global cyclic ranking consistency on cross-camera tracklets (Sec. 3.2). Best viewed in colour.

3 Deep Association Learning

Approach Overview. Our goal is to learn a re-id matching model to discriminate the appearance difference and reliably associate the video tracklets across disjoint camera views without utilising any ID labels. Towards this goal, we propose a novel *Deep Association Learning* (DAL) scheme that optimises a deep CNN model based on the learning objective derived based on two types of consistency. As illustrated in Figure 2, we explore the *local space-time consistency* and *global cyclic ranking consistency* to formulate two top-push margin-based association losses. In particular, two sets of "anchors" are gradually learned all along the training process for our loss formulation. They are **(1)** a set of *intra-camera anchors* $\{x_{k,i}\}_{i=1}^{N_k}$ that denote the intra-camera feature representations of N_k tracklets under camera k; and **(2)** a set of *cross-camera anchors* $\{a_{k,i}\}_{i=1}^{N_k}$, with each representing the cross-camera feature representation merged by the intra-camera feature representations of two highly associated tracklets from disjoint camera views. Overall, the DAL scheme consists of two batch-wise iterative procedures: **(a)** intra-camera association learning and **(b)** cross-camera association learning, as elaborated in the following.

3.1 Intra-Camera Association Learning

Intra-camera association learning aims at discriminating intra-camera video tracklets. To this end, we formulate a top-push margin-based intra-camera association loss in the form of the hinge loss based on the ranking relationship of each image frame in association to all the video tracklets from the same camera view. This loss is formulated in three steps as follows.

(1) Learning Intra-Camera Anchors. On account of the *local space-time consistency* as depicted Figure 1, each video tracklet can simply be represented as a univocal sequence-level feature representation by utilising certain temporal pooling strategy, such as max-pooling or mean-pooling [28, 45]. This, however, is time-consuming to compute at each mini-batch learning iteration, as it requires to feed-forward all image frames of each video tracklet through the deep model. To overcome this problem, we propose to represent a tracklet from camera k as an *intra-camera anchor* $x_{k,i}$, which is the intra-camera tracklet representation incrementally updated by the frame representation $f_{k,p}$ of any constituent image frame from the same source tracklet all through the training process. Specifically, the exponential moving average (EMA) strategy is adopted to update each anchor $x_{k,i}$ as follows.

$$x_{k,i}^{t+1} \leftarrow x_{k,i}^{t} - \eta\left(\ell_2(x_{k,i}^{t}) - \ell_2(f_{k,p}^{t})\right), \text{ if } i = p \tag{1}$$

where η refers to the update rate (set to 0.5), $\ell_2(\cdot)$ is ℓ_2 normalisation (i.e. $\|\ell_2(\cdot)\|_2 = 1$), and t is the mini-batch learning iteration. As $x_{k,i}$ is initialised as the mean of the frame representations for each tracklet and incrementally updated as Eq. (1), the intra-camera anchor is consistently learned all along with the model learning progress to represent each tracklet.

(2) Tracklet Association Ranking. Given the set of incrementally updated *intra-camera anchors* $\{x_{k,i}\}_{i=1}^{N_k}$ for camera k, the ranking relationship of the frame representation $f_{k,p}$ in association to all intra-camera anchors from the same camera k can be generated based on pairwise similarity measure. We use the ℓ_2 distance to measure the pairwise similarities between an in-batch frame representation $f_{k,p}$ and all the intra-camera anchor $\{x_{k,i}\}_{i=1}^{N_k}$. Accordingly, a ranking list is obtained by sorting the pairwise similarities of $f_{k,p}$ w.r.t. $\{x_{k,i}\}_{i=1}^{N_k}$, with the rank-1 (top-1) intra-camera anchor having the minimal pairwise distance:

$$\{D_{p,i}|D_{p,i} = \left\|\ell_2(f_{k,p}) - \ell_2(x_{k,i})\right\|_2,\ i \in N_k\} \xrightarrow{\text{ranking}} D_{p,t} = \min_{i\in[1,N_k]} D_{p,i} \tag{2}$$

where $\{D_{p,i}\}_{i=1}^{N_k}$ is the set of pairwise distances between $f_{k,p}$ and $\{x_{k,i}\}_{i=1}^{N_k}$; while $D_{p,t}$ denotes the pairwise distance between $f_{k,p}$ and the rank-1 tracklet $x_{k,t}$.

(3) Intra-Camera Association Loss. Given the ranking list for the frame representation $f_{k,p}$ (Eq. (2)), the intra-camera rank-1 tracklet $x_{k,t}$ should ideally correspond to the source tracklet $x_{k,p}$ that contains the same constituent frame due to the *local space-time consistency*. We therefore define a top-push margin-based intra-camera association loss to enforce proper association of each frame to the source tracklet for discriminative model learning:

$$\mathcal{L}_I = \begin{cases} [D_{p,p} - D_{p,t} + m]_+, & \text{if } p \neq t \text{ (The rank-1 is not the source tracklet)} \\ [D_{p,p} - \overline{D_{j,t}} + m]_+, & \text{if } p = t \text{ (The rank-1 is the source tracklet)} \end{cases} \tag{3}$$

where $[\cdot]_+ = \max(0, \cdot)$, $D_{p,p}$ is the pairwise distance between $f_{k,p}$ and $x_{k,p}$ (the source tracklet), $\overline{D_{j,t}} = \frac{1}{M}\sum_{j=1}^{M} D_{j,t}$ is the averaged rank-1 pairwise distance of the M sampled image frames from camera k in a mini-batch. m is the margin that enforces the deep model to assign the source tracklet as the top-rank. More specifically, if the rank-1 is not the source tracklet (i.e. $p \neq t$), $\mathcal{L}_I$ will correct the model by imposing a large penalty to push the source tracklet to the top-rank. Otherwise, $\mathcal{L}_I$ will further minimise the intra-tracklet variation w.r.t. the averaged rank-1 pairwise distance in each mini-batch. Since $\mathcal{L}_I$ is computed based on the sampled image frames and the up-to-date intra-camera anchors in each mini-batch, it can be efficiently optimised by the standard stochastic gradient descent to adjust the deep CNN parameters iteratively. Overall, $\mathcal{L}_I$ encourages to learn the discrimination on intra-camera tracklets for facilitating the more challenging cross-camera association, as described next.

3.2 Cross-Camera Association Learning

A key of video re-id is to leverage the cross-camera ID pairing information for model learning. However, such information is missing in unsupervised learning. We overcome this problem by self-discovering the cross-camera tracklet association in a progressive way during model training. To permit learning expressive representation invariant to the cross-camera appearance variations inherently carried in associated tracklet pairs from disjoint camera views, we formulate another top-push margin-based intra-camera association loss in the same form as Eq. (3). Crucially, we extend the tracklet representation to carry the information of cross-camera appearance variations by incrementally learning a set of *cross-camera anchors*. This intra-camera association loss is formulated in three steps as below.

(1) Cyclic Ranking. Given the incrementally updated intra-camera anchors (Eq. (1)), we propose to exploit the underlying relations between tracklets for discovering the association between tracklets across different cameras. Specifically, a cyclic ranking process is conducted to attain the pair of highly associated intra-camera anchors across cameras as follows.

$$x_{k,i} \xrightarrow{\text{ranking in cam } l} Dc_{p,t} = min_{i\in[1,N_l]} \xrightarrow{\text{ranking back in cam } k} Dc_{q,j} = min_{i\in[1,N_k]} \\ \text{Global Cyclic Ranking Consistency: } j=i \tag{4}$$

where $Dc_{p,t}$ denotes the cross-camera pairwise distance between two intra-camera anchors: $x_{k,p}$ from camera k and $x_{l,t}$ from another camera l. Both $Dc_{p,t}$ and $Dc_{q,j}$ denote the rank-1 pairwise distance. The pairwise distance and the ranking are computed same as Eq. (2). With Eq. (4), we aim to discover the most associated intra-camera anchors across cameras under the criterion of *global cyclic ranking consistency*: $x_{k,p}$ and $x_{l,t}$ are mutually the rank-1 match pair to each other when one is given as a query to search for the best-matched intra-camera anchor in the other camera view. This cyclic ranking process is conceptually related to the cycle-consistency constraints formulated to enforce the pairwise correspondence between similar instances [12, 31, 49]. In particular, our *global cyclic ranking consistency* in this process aims to exploit the mutual consistency induced by transitivity for discovering the highly associated tracklets across disjoint camera views all along the model training process.

(2) Learning Cross-Camera Anchors. Based on *global cyclic ranking consistency*, we define the cross-camera representation as a *cross-camera anchor* $a_{k,i}$ by merging two highly associated intra-camera anchors as depicted in Figure 2 and detailed below.

$$a_{k,i}^{t+1} \leftarrow \begin{cases} \frac{1}{2}\left(\ell_2(x_{k,i}^{t+1}) + \ell_2(x_{l,t}^{t})\right), & \text{if } j=i \quad \text{(Cyclic ranking consistent)} \\ x_{k,i}^{t+1}, & \text{others} \end{cases} \tag{5}$$

where $a_{k,i}$ is simply a counterpart of $x_{k,i}$. Each cross-camera anchor is updated as the arithmetic mean of two intra-camera anchors if the consistency condition is fulfilled (i.e. $j=i$), otherwise as the same intra-camera anchor. As the deep model is updated continuously to discriminate the appearance difference among tracklets, more intra-camera anchors are progressively discovered to be highly associated. That is, all along the training process, more cross-camera anchors are gradually updated by merging the highly associated intra-camera anchors to carry the information of cross-camera appearance variations induced by the tracklet pairs that come from disjoint camera views but potentially depict the same identities.

(3) Cross-Camera Association Loss. Given the continuously updated *cross-camera anchors* $\{a_{k,i}\}_{i=1}^{N_k}$, we define another top-push margin-based cross-camera association loss in the same form as Eq. (3) to enable learning from cross-camera appearance variations:

$$\mathcal{L}_C = \begin{cases} [Da_{p,p} - D_{p,t} + m]_+, & \text{if } p \neq t \quad \text{(The rank-1 is not the source tracklet)} \\ [Da_{p,p} - \overline{D_{j,t}} + m]_+, & \text{if } p = t \quad \text{(The rank-1 is the source tracklet)} \end{cases} \tag{6}$$

where $Da_{p,p}$ denotes the pairwise distance between the frame representation $f_{k,p}$ and the cross-camera anchor $a_{k,p}$. Both $D_{p,t}$ and $\overline{D_{j,t}}$ are the same quantities as $\mathcal{L}_I$ in Eq. (3). As depicted in Figure 2 and in the same spirit as $\mathcal{L}_I$, the cross-camera association loss $\mathcal{L}_C$ enforces the deep model to push the best-associated cross-camera anchor as the top-rank, so as to align the frame representation $f_{k,p}$ towards the corresponding cross-camera representation.

3.3 Model Training

Overall Learning Objective. The final learning objective for DAL is to jointly optimise two association losses (Eq. (3), (6)) as follows.

$$\mathcal{L}_{DAL} = \mathcal{L}_I + \lambda \mathcal{L}_C \tag{7}$$

where λ is a tradeoff parameter that is set to 1 to ensure both loss terms contribute equally to the learning process. The margin m in both Eq. (3) and Eq. (6) is empirically set to 0.2 in our experiments. The algorithmic overview of model training is summarised in Algorithm 1.

Complexity Analysis. We analyse the per-batch per-sample complexity cost induced by DAL. In association ranking (Eq. (2)), the pairwise distances are computed between each in-batch image frame and N_k intra-camera anchors for each camera, which leads to a computation complexity of $\mathcal{O}(N_k)$ for distance computation and $\mathcal{O}(N_k log(N_k))$ for ranking. Similarly, in cyclic ranking (Eq. (4)), the total computation complexity is $\mathcal{O}(N_l + N_k) + \mathcal{O}(N_l log(N_l) + N_k log(N_k))$. All the distance measures are simply computed by matrix manipulation on GPU with single floating point precision for computational efficiency.

Algorithm 1 Deep Association Learning.

Input: Unlabelled video tracklets captured from different cameras.
Output: A deep CNN model for re-id matching.
for $t = 1$ **to** *max_iter* **do**
  Randomly sample a mini-batch of image frames.
  Network forward propagation.
  Tracklet association ranking on the *intra-camera anchors* (Eq. (2)).
  Compute two margin-based association loss terms (Eq. (3), (6)).
  Update the corresponding *intra-camera anchors* based on the EMA strategy (Eq. (1)).
  Update the corresponding *cross-camera anchors* based on cyclic ranking (Eq. (4), (5)).
  Network update by back-propagation (Eq. (7)).
end for

4 Experiments

4.1 Evaluation on Unsupervised Video Person Re-ID

Datasets. We conduct extensive experiments on three video person re-id benchmark datasets, including PRID 2011 [16], iLIDS-VID [37] and MARS [45] (Figure 3). The PRID 2011 dataset contains 1,134 tracklets captured from two disjoint surveillance cameras with 385 and 749 tracklets from the first and second cameras. Among all video tracklets, 200 persons are captured in both cameras. The iLIDS-VID dataset includes 600 video tracklets of 300 persons. Each person has 2 tracklets from two non-overlapping camera views in an airport arrival hall. The MARS has a total of 20,478 tracklets of 1,261 persons captured from a camera network with 6 near-synchronized cameras at a university campus. All the tracklets were automatically generated by the DPM detector [11] and the GMMCP tracker [8].

Evaluation Protocols. For PRID 2011, following [26, 37, 42] we use the tracklet pairs from 178 persons, with each tracklet containing over 27 frames. These 178 persons are further randomly divided into two halves (89/89) for training and testing. For iLIDS-VID, all 300 persons are also divided into two halves (150/150) for training and testing. For both datasets,

(a) PRID 2011 (b) iLIDS-VID (c) MARS

Figure 3: Example pairs of tracklets from three benchmark datasets. Cross-camera variations include changes in illumination, viewpoints, resolution, occlusion, background clutter, human poses, etc.

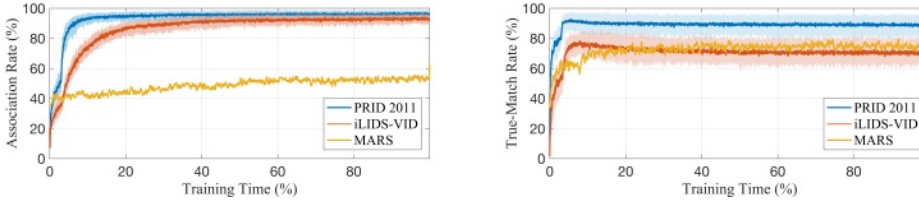
Datasets	PRID 2011				iLIDS-VID				MARS				
Rank@k	1	5	10	20	1	5	10	20	1	5	10	20	mAP
DVDL [18]	40.6	69.7	77.8	85.6	25.9	48.2	57.3	68.9	-	-	-	-	-
STFV3D [25]	42.1	71.9	84.4	91.6	37.0	64.3	77.0	86.9	-	-	-	-	-
MDTS-DTW [27]	41.7	67.1	79.4	90.1	31.5	62.1	72.8	82.4	-	-	-	-	-
UnKISS [19]	59.2	81.7	90.6	96.1	38.2	65.7	75.9	84.1	-	-	-	-	-
DGM+IDE [42]	56.4	81.3	88.0	96.4	36.2	62.8	73.6	82.7	36.8	54.0	61.6	68.5	21.3
Stepwise [26]	80.9	95.6	98.8	99.4	41.7	66.3	74.1	80.7	23.6	35.8	-	44.9	10.5
DAL (ResNet50)	85.3	97.0	98.8	99.6	56.9	80.6	87.3	91.9	46.8	63.9	71.6	77.5	21.4
DAL (MobileNet)	84.6	96.3	98.4	99.1	52.8	76.7	83.4	91.6	49.3	65.9	72.2	77.9	23.0

Table 1: Evaluation on three benchmarks in comparison to the state-of-the-art unsupervised video re-id methods. Red: the best performance. Blue: the second best performance. '-': no reported results.

we repeat 10 random training/testing ID splits as [37] to ensure statistically stable results. The average Cumulated Matching Characteristics (CMC) are adopted as the performance metrics. For MARS, we follow the standard training/testing split [45]: all tracklets of 625 persons for training and the remaining tracklets of 636 persons for testing. Both the averaged CMC and the mean Average Precision (mAP) are used to measure re-id performance on MARS. Note, our method does not utilise any ID labels for model initialisation or training.

Implementation Details. We implement our DAL scheme in Tensorflow [1]. To evaluate its generalisation ability of incorporating with different network architectures, we adopt two standard CNNs as the backbone networks: ResNet50 [14] and MobileNet [17]. Both deep models are initialised with weights pre-trained on ImageNet [9]. On the small-scale datasets (PRID 2011 and iLIDS-VID), we apply the RMSProp optimiser [33] to train the DAL for 2×10^4 iterations, with an initial learning rate of 0.045 and decayed exponentially by 0.94 every 2 epochs. On the large-scale dataset (MARS), we adopt the standard stochastic gradient descent (SGD) to train the DAL for 1×10^5 iterations, with an initial learning rate of 0.01 and decayed to 0.001 in the last 5×10^4 iterations. The batch size is all set to 64. At test time, we obtain the tracklet representation by max-pooling on the image frame features followed by ℓ_2 normalisation. We compute the ℓ_2-distance between the cross-camera tracklet representations as the similarity measure for the final video re-id matching.

Comparison to the state-of-the-art methods. We compare DAL against six state-of-the-art video-based unsupervised re-id methods: DVDL [18], STFV3D [25], MDTS-DTW [27], UnKISS [19], DGM+IDE [42], and Stepwise [26]. Among all methods, DAL is the only unsupervised deep re-id model that is optimised in an end-to-end manner. Table 1 shows a clear performance superiority of DAL over all other competitors on the three benchmark datasets. In particular, the rank-1 matching accuracy is improved by 4.4%(85.3-80.9) on PRID 2011, 15.2%(56.9-41.7) on iLIDS-VID and 12.5%(49.3-36.8) on MARS. This consistently shows the advantage of DAL over existing methods for unsupervised video re-id due to the joint effect of optimising two association losses to enable learning feature representation invariant to cross-camera appearance variations whilst discriminative to appearance difference. Note,



(a) Evolution on association rate. (b) Evolution on true-match rate.

Figure 4: Evolution on cross-camera tracklet association. The shaded areas denote the varying range of 10-split results repeated on PRID 2011 and iLIDS-VID. Best viewed in colour.

the strongest existing model DGM+IDE [42] additionally uses ID label information from one camera view for model initialisation, whilst Stepwise [26] assumes one tracklet per ID per camera by implicitly using ID labels. In contrast, DAL uses neither of such additional label information for model initialisation or training. More crucially, DAL consistently produces similar strong re-id performance with different network architectures (ResNet50 and MobileNet), which demonstrates its applicability to existing standard CNNs.

4.2 Component Analyses and Further Discussions

Effectiveness of two association losses. The DAL trains the deep CNN model based on the joint effect of two association losses: (1) intra-camera association loss $\mathcal{L}_I$ (Eq. (3)) and (2) cross-camera association loss $\mathcal{L}_C$ (Eq. (3)). We evaluate the individual effect of each loss term by eliminating the other term from the overall learning objective (Eq. (7)). As shown in Table 2, jointly optimising two losses leads to the best model performance. This indicates the complementary benefits of the two loss terms in discriminative feature learning. Moreover, applying $\mathcal{L}_C$ alone has already achieved better performance as compared to the state-of-the-art methods in Table 1. When comparing with $\mathcal{L}_I$+$\mathcal{L}_C$, applying $\mathcal{L}_C$ alone only drop the rank-1 accuracy by 3.0%(84.6-81.6), 5.4%(52.8-47.4), 1.2%(49.3-48.1) on PRID 2011, iLIDS-VID, MARS respectively. This shows that even optimising the cross-camera association loss *alone* can still yield competitive re-id performance, which owes to its additional effect in enhancing cross-camera invariant representation learning by reliably associating tracklets across disjoint camera views all along the training process.

Evolution of cross-camera tracklet association. As aforementioned, learning representation robust to cross-camera variations is a key to learning an effective video re-id model. To understand the effect of utilising the cyclic ranking consistency to discover highly associated tracklets during training, we track the proportion of *cross-camera anchors* that are updated to denote the cross-camera representation by merging two highly associated tracklets (*intra-camera anchors*). Figure 4(a) shows that on PRID 2011 and iLIDS-VID, 90+% tracklets find their highly associated tracklets under another camera at the end of training. On the much noisier large-scale MARS dataset, the DAL can still associate more than half of

Datasets	PRID 2011				iLIDS-VID				MARS				
Rank@k	1	5	10	20	1	5	10	20	1	5	10	20	mAP
$\mathcal{L}_I$ Only	62.7	85.7	92.1	96.7	31.7	55.2	67.5	78.6	41.6	59.0	66.2	73.2	16.8
$\mathcal{L}_C$ Only	81.6	95.2	98.1	99.7	47.4	72.6	81.5	89.2	48.1	65.3	71.4	77.6	22.6
$\mathcal{L}_I$+$\mathcal{L}_C$	84.6	96.3	98.4	99.1	52.8	76.7	83.4	91.6	49.3	65.9	72.2	77.9	23.0

Table 2: Effectiveness of two association losses. Red: the best performance. CNN: MobileNet.

Datasets	PRID 2011				iLIDS-VID				MARS				
Rank@k	1	5	10	20	1	5	10	20	1	5	10	20	mAP
DAL ($\mathcal{L}_I$+$\mathcal{L}_C$)	84.6	96.3	98.4	99.1	52.8	76.7	83.4	91.6	49.3	65.9	72.2	77.9	23.0
ID-Supervised	84.3	98.1	99.2	99.8	51.5	76.0	83.8	89.9	71.8	86.8	90.7	93.3	51.5

Table 3: Comparison with supervised counterparts. Red: the best performance. CNN: MobileNet.

tracklets (>50%) across cameras. Importantly, as seen in Figure 4(b), among self-discovered associated cross-camera tracklet pairs, the percentage of true-match pairs at the end of training is approximately 90% on PRID 2011, 75% on iLIDS-VID, and 77% on MARS, respectively. This shows compellingly the strong capability of DAL in self-discovering the unknown cross-camera tracklet associations without learning from manually labelled data.

Comparison with supervised counterparts. We further compare DAL against the supervised counterpart trained using ID labelled data with the identical CNN architecture (MobileNet), denoted as ID-Supervised. This ID-Supervised is trained by the cross-entropy loss computed on the ID labels. Results in Table 3 show that: (1) On PRID 2011 and iLIDS-VID, DAL performs similarly well as the ID-Supervised. This is highly consistent with our observations of high tracklet association rate in in Figure 4, indicating that discovering more cross-camera highly associated tracklets can help to learn a more discriminative re-id model that is robust to cross-camera variations. (2) On MARS, there is a clear performance gap between the supervised and unsupervised models. This is largely due to a relatively low tracklet association rate arising from the difficulty of discovering cross-camera tracklet associations in a larger identity population among much noisier tracklets, as indicated in Figure 4(a).

5 Conclusions

In this work, we present a novel *Deep Association Learning* (DAL) scheme for unsupervised video person re-id using unlabelled video tracklets extracted from surveillance video data. Our DAL permits deep re-id models to be trained without any ID labelling for training data, which is therefore more scalable to deployment on large-sized surveillance video data than supervised learning based models. In contrast to existing unsupervised video re-id methods that either require more stringent one-camera ID labelling or per-camera tracklet filtering, DAL is capable of learning to automatically discover the more reliable cross-camera tracklet associations for addressing the video re-id task without utilising ID labels. This is achieved by jointly optimising two margin-based association losses formulated based on the *local space-time consistency* and *global cyclic ranking consistency*. Extensive comparative experiments on three video person re-id benchmarks show compellingly the clear advantages of the proposed DAL scheme over a wide variety of state-of-the-art unsupervised video person re-id methods. We also provide detailed component analyses to further discuss the insights on how each part of our method design contributes towards the overall model performance.

Acknowledgements

This work was partly supported by the China Scholarship Council, Vision Semantics Limited, the 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] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016.

[2] Ejaz Ahmed, Michael Jones, and Tim K Marks. An improved deep learning architecture for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[3] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.

[4] Shi-Zhe Chen, Chun-Chao Guo, and Jian-Huang Lai. Deep ranking for person re-identification via joint representation learning. *IEEE Transactions on Image Processing*, 2016.

[5] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[6] Yanbei Chen, Xiatian Zhu, Shaogang Gong, et al. Person re-identification by deep learning multi-scale representations. In *Workshop of IEEE International Conference on Computer Vision*, 2017.

[7] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[8] Afshin Dehghan, Shayan Modiri Assari, and Mubarak Shah. Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[10] Shengyong Ding, Liang Lin, Guangrun Wang, and Hongyang Chao. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 2015.

[11] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010.

[12] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[13] Shaogang Gong, Marco Cristani, Shuicheng Yan, and Chen Change Loy. *Person re-identification*. Springer, 2014.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[15] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[16] Martin Hirzer, Csaba Beleznai, Peter M Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *Scandinavian Conference on Image Analysis*, 2011.

[17] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[18] Srikrishna Karanam, Yang Li, and Richard J Radke. Person re-identification with discriminatively trained viewpoint invariant dictionaries. In *IEEE International Conference on Computer Vision*, 2015.

[19] Furqan M Khan and Francois Bremond. Unsupervised data association for metric learning in the context of multi-shot person re-identification. In *IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2016.

[20] Shuang Li, Slawomir Bak, Peter Carr, and Xiaogang Wang. Diversity regularized spatiotemporal attention for video-based person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[21] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[22] Wei Li, Xiatian Zhu, and Shaogang Gong. Person re-identification by deep joint learning of multi-loss classification. In *International Joint Conference of Artificial Intelligence*, 2017.

[23] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[24] Jiawei Liu, Zheng-Jun Zha, QI Tian, Dong Liu, Ting Yao, Qiang Ling, and Tao Mei. Multi-scale triplet cnn for person re-identification. In *ACM International Conference on Multimedia*, 2016.

[25] Kan Liu, Bingpeng Ma, Wei Zhang, and Rui Huang. A spatio-temporal appearance representation for video-based pedestrian re-identification. In *IEEE International Conference on Computer Vision*, 2015.

[26] Zimo Liu, Dong Wang, and Huchuan Lu. Stepwise metric promotion for unsupervised video person re-identification. In *IEEE International Conference on Computer Vision*, 2017.

[27] Xiaolong Ma, Xiatian Zhu, Shaogang Gong, Xudong Xie, Jianming Hu, Kin-Man Lam, and Yisheng Zhong. Person re-identification by unsupervised video matching. *Pattern Recognition*, 2017.

[28] Niall McLaughlin, Jesus Martinez del Rincon, and Paul Miller. Recurrent convolutional network for video-based person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[29] Sakrapee Paisitkriangkrai, Chunhua Shen, and Anton van den Hengel. Learning to rank in person re-identification with metric ensembles. In *IEEE Conference on Computer Vision and Pattern Recognition*, Boston, Massachusetts, USA, 2015.

[30] B. Prosser, W-S. Zheng, S. Gong, and T. Xiang. Person re-identification by support vector ranking. In *British Machine Vision Conference*, Aberystwyth, Wales, September 2010.

[31] Ozan Sener, Hyun Oh Song, Ashutosh Saxena, and Silvio Savarese. Learning transferrable representations for unsupervised domain adaptation. In *Advances in Neural Information Processing Systems*, 2016.

[32] Yifan Sun, Liang Zheng, Weijian Deng, and Shengjin Wang. Svdnet for pedestrian retrieval. In *IEEE International Conference on Computer Vision*, 2017.

[33] Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, pages 26–31, 2012.

[34] Faqiang Wang, Wangmeng Zuo, Liang Lin, David Zhang, and Lei Zhang. Joint learning of single-image and cross-image representations for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[35] Hanxiao Wang, Xiatian Zhu, Shaogang Gong, and Tao Xiang. Person re-identification in identity regression space. *International Journal of Computer Vision*, 2018.

[36] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[37] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by video ranking. In *European Conference on Computer Vision*, 2014.

[38] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by discriminative selection in video ranking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016.

[39] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2016.

[40] Shuangjie Xu, Yu Cheng, Kang Gu, Yang Yang, Shiyu Chang, and Pan Zhou. Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[41] Yichao Yan, Bingbing Ni, Zhichao Song, Chao Ma, Yan Yan, and Xiaokang Yang. Person re-identification via recurrent feature aggregation. In *European Conference on Computer Vision*, 2016.

[42] Mang Ye, Andy J Ma, Liang Zheng, Jiawei Li, and Pong C Yuen. Dynamic label graph matching for unsupervised video re-identification. In *IEEE International Conference on Computer Vision*, 2017.

[43] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Deep metric learning for person re-identification. In *IEEE International Conference on Pattern Recognition*, 2014.

[44] Jinjie You, Ancong Wu, Xiang Li, and Wei-Shi Zheng. Top-push video-based person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[45] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. Mars: A video benchmark for large-scale person re-identification. In *European Conference on Computer Vision*, 2016.

[46] W-S. Zheng, S. Gong, and T. Xiang. Re-identification by relative distance comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.

[47] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *IEEE International Conference on Computer Vision*, 2017.

[48] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. Camera style adaptation for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[49] Tinghui Zhou, Philipp Krahenbuhl, Mathieu Aubry, Qixing Huang, and Alexei A Efros. Learning dense correspondence via 3d-guided cycle consistency. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[50] Zhen Zhou, Yan Huang, Wei Wang, Liang Wang, and Tieniu Tan. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[51] Xiaoke Zhu, Xiao-Yuan Jing, Fei Wu, and Hui Feng. Video-based person re-identification by simultaneously learning intra-video and inter-video distance metrics. In *International Joint Conference of Artificial Intelligence*, 2016.

[52] Xiatian Zhu, Botong Wu, Dongcheng Huang, and Wei-Shi Zheng. Fast openworld person re-identification. *IEEE Transactions on Image Processing*, 2017.