# Tracklet Self-Supervised Learning for Unsupervised Person Re-Identification

**Guile Wu,[1] Xiatian Zhu,[2] Shaogang Gong[1]**
[1]Queen Mary University of London, [2]Vision Semantics Limited
guile.wu@qmul.ac.uk, eddy.zhuxt@gmail.com, s.gong@qmul.ac.uk

## Abstract

Existing unsupervised person re-identification (re-id) methods mainly focus on cross-domain adaptation or one-shot learning. Although they are more scalable than the supervised learning counterparts, relying on a relevant labelled source domain or one labelled tracklet per person initialisation still restricts their scalability in real-world deployments. To alleviate these problems, some recent studies develop unsupervised tracklet association and bottom-up image clustering methods, but they still rely on explicit camera annotation or merely utilise suboptimal global clustering. In this work, we formulate a novel tracklet self-supervised learning (TSSL) method, which is capable of capitalising directly from abundant unlabelled tracklet data, to optimise a feature embedding space for both video and image unsupervised re-id. This is achieved by designing a comprehensive unsupervised learning objective that accounts for tracklet frame coherence, tracklet neighbourhood compactness, and tracklet cluster structure in a unified formulation. As a pure unsupervised learning re-id model, TSSL is end-to-end trainable at the absence of source data annotation, person identity labels, and camera prior knowledge. Extensive experiments demonstrate the superiority of TSSL over a wide variety of the state-of-the-art alternative methods on four large-scale person re-id benchmarks, including Market-1501, DukeMTMC-ReID, MARS and DukeMTMC-VideoReID.

## Introduction

The key in person re-identification (re-id) is learning a discriminative feature representation model (Liu et al. 2019b; Zhang et al. 2019b; Tesfaye et al. 2019; Dong, Gong, and Zhu 2019). While existing *supervised learning* based re-id methods have advanced significantly (Fu et al. 2019; Wu, Zhu, and Gong 2019b), they fundamentally suffer from an *unrealistic* assumption of requiring a large set of cross-camera labelled training data (Yu et al. 2019; Li, Zhu, and Gong 2018a). To address this, recent studies have shifted to capitalise abundant unlabelled data for *unsupervised* model optimisation (Lin et al. 2019; Zhang et al. 2019a).

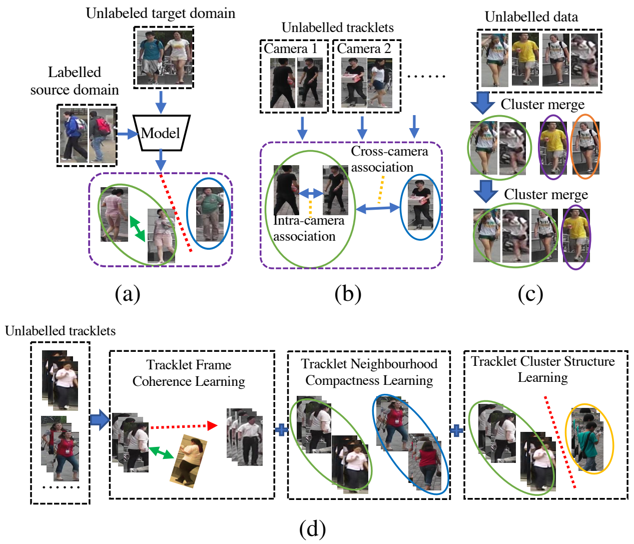

Figure 1: Illustration of four unsupervised person re-id learning strategies that aim at minimising the training data annotation efforts. (a) Unsupervised cross-domain adaptation. (b) Unsupervised tracklet association. (c) Bottom-up image clustering. (d) Tracklet self-supervised learning (*proposed*). Our method eliminates the need for both person identity labels and camera view labels for pure unsupervised video and image person re-id.

Whilst hand-crafted descriptors (Liao et al. 2015) can be used for re-id matching without label supervision, they often yield rather poor performance (Wang et al. 2018). A more effective approach to unsupervised re-id is by cross-domain transfer learning (Liu et al. 2019a; Zhong et al. 2018; Yu et al. 2019). These methods typically pre-train the model in a labelled source domain and then adapt it to an unlabelled target domain (Fig. 1(a)). This assumes sufficient knowledge overlap between the source and target domains, which nonetheless is not always valid in practice. Alternatively, some methods explore one-shot training per identity in the target domain (Wu et al. 2018a; Ye et al. 2017; Liu, Wang, and Lu 2017). This reduces the labelling efforts but remains unscalable in real-world applications. A scalable approach is by *pure unsupervised learning*, e.g. unsupervised tracklet association (Li, Zhu, and Gong 2018a; Chen, Zhu, and Gong 2018) and global bottom-up image clustering (Lin et al. 2019). They require *no* person identity labelling in any domain; However, the former assumes the availability of camera annotations (Fig. 1(b)), while the latter merely utilises suboptimal global clustering (Fig. 1(c)).

In this work, we investigate pure unsupervised person re-identification, where neither labelled source data nor initial tracklet identity label are available. To this end, we propose *tracklet self-supervised learning* (TSSL) to optimise a feature embedding space for both video and image unsupervised re-id. TSSL makes a good use of the intrinsic tracklet structure and appearance information, eliminating the notorious need for both person identity and camera labels (Fig. 1(d)). Specifically, we formulate a comprehensive tracklet self-supervised learning objective, covering three self-supervision mining: tracklet frame coherence learning, tracklet neighbourhood compactness learning, and tracklet cluster structure learning. These learning components are derived at different granularities, ranging from per-tracklet and local tracklet neighbourhoods to global tracklet clusters. Consequently, they present high complementary interaction when integrated into a unified learning objective function. The ultimate objective is to train a feature embedding model discriminative for person re-id matching.

The **contributions** of this work are: **(I)** We propose an idea of tracklet self-supervised learning for unsupervised person re-identification. This eliminates the need for person identity labels and camera view annotation simultaneously, enabling both highly scalable video and image re-id deployments in real-world applications. **(II)** We formulate a novel tracklet self-supervised learning (TSSL) method based on comprehensive self-supervision mining on unlabelled tracklet data from individual tracklets to tracklet clusters. As a unified learning architecture, TSSL is end-to-end trainable. **(III)** Given that the proposed TSSL does not use any label information, a direct comparison between TSSL and the state-of-the-art alternative methods (*e.g.* cross-domain re-id (Liu et al. 2019a; Zhong et al. 2019)) might not be fair, but extensive experiments still demonstrate the superiority of TSSL against the state-of-the-arts on two large-scale image benchmarks (Market-1501 (Zheng et al. 2015) and DukeMTMC-ReID (Zheng, Zheng, and Yang 2017)) and two large-scale video benchmarks (MARS (Zheng et al. 2016) and DukeMTMC-VideoReID (Wu et al. 2018a)).

## Related Work

Most existing person re-id methods are based on supervised learning, which require labelled pairs of person images for training (Li, Zhu, and Gong 2018b; Wu, Zhu, and Gong 2019a), leading to limited scalability in deployment. In contrast, unsupervised re-id is capable of learning from unlabelled data without exhaustive manual annotation, allowing to leverage massive available unlabelled data. In this section, we mainly review and discuss unsupervised person re-id.

### Unsupervised Cross-Domain Person Re-ID

Transfer learning is one of the most important strategies for addressing unsupervised re-id, *i.e.* unsupervised cross-domain person re-id. Existing methods typically pre-train a model in source domains with rich labelled training data, and then transfer this model to an unlabelled target domain (Yu et al. 2019; Liu et al. 2019a; Zhong et al. 2018; 2019). In (Yu et al. 2019), Yu et al. propose using a set of labelled persons from a source domain as the references to facilitate soft multi-label estimation for unlabelled persons in a target domain. In (Yang et al. 2019), Yang et al. introduce a patch-based model to learn discriminative features. They pre-train this model in a large-scale labelled source dataset before fine-tuning it in an unlabelled target dataset based on patch-level and image-level learning constraints. In (Wang et al. 2018), Wang et al. transfer both identity and attribute information from a labelled source domain to an unlabelled target domain. This is achieved by extracting attribute-semantic and identity-discriminative feature representations. Different from these methods, we focus on *pure* unsupervised re-id, where no prior knowledge is available from any labelled source domain. This is for further scaling the learning algorithm to arbitrarily unconstrained and unlabelled domains, without the need for selecting relevant source domains.

### One-Shot Person Re-ID

One-shot person re-id is a recently developed technique for training data annotation minimisation. It often assumes long tracklets and/or the spatio-temporal topology knowledge for obtaining automatically person identity labels (Wu et al. 2018a; Ye, Lan, and Yuen 2018). Specifically, Liu, Wang, and Lu (Liu, Wang, and Lu 2017) perform reciprocal nearest neighbour search for negative sample mining to realise unsupervised video re-id. In (Ye et al. 2017), Ye et al. propose a dynamic graph matching method to iteratively update the model based on intermediately estimated labels. In (Wu et al. 2018a), Wu et al. gradually exploit unlabelled tracklets with reliable pseudo labels for online model update. While dropping the dependence on a labelled source domain, they still require one labelled tracklet per identity in the target domain for model initialisation. On the contrary, our TSSL method exploits unlabelled tracklet data alone with no need of labelled source data or one-shot tracklet annotation.

### Pure Unsupervised Person Re-ID

There are only a few existing studies focus on *pure* unsupervised re-id without using any person identity annotation. Li, Zhu, and Gong (Li, Zhu, and Gong 2018a) conduct tracklet association learning within-camera and cross-camera concurrently. Chen, Zhu, and Gong (Chen, Zhu, and Gong 2018) leverage both intra-camera and cross-camera anchors to improve tracklet association learning. But these methods assume the availability of camera view annotation, limiting their usability when no camera information is given. This problem is resolved by bottom-up clustering re-id (Lin et al. 2019) wherein a CNN model is trained from unlabelled target data alone. Only focusing on cluster-level repelling and merging, this method is suboptimal due to ignoring the latent variational information of each sample. Unlike these methods, we model the self-supervision process on unlabelled tracklets from per-tracklet and small neighbourhoods

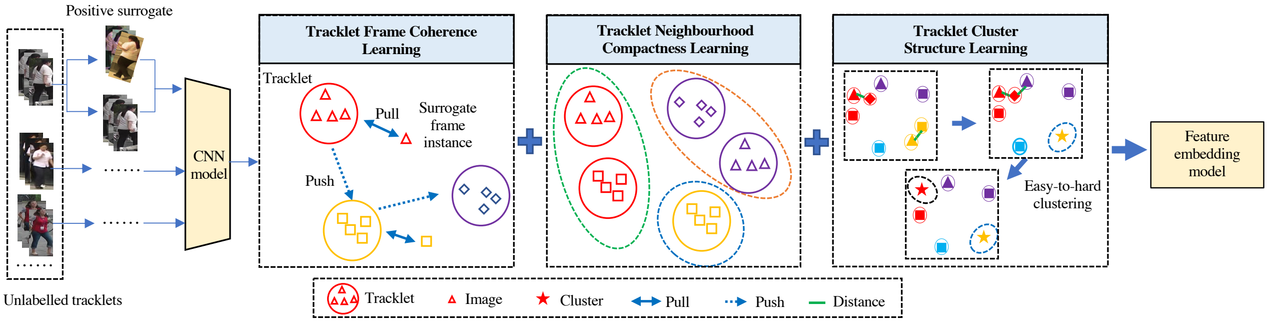

Figure 2: Overview of Tracklet Self-Supervised Learning (TSSL). TSSL aims to learn a feature embedding model from unlabelled training tracklets enabling both video and image re-id. To this end, TSSL trains the model with three self-supervised optimisation constraints: (1) tracklet frame coherence learning, (2) tracklet neighbourhood compactness learning, and (3) tracklet cluster structure learning. These constrained are integrated in an end-to-end pipeline, designed specially for mining the intrinsic tracklet structural discrimination information from the unlabelled data at different granularities in a progressively learning manner.

to global cluster structure in a unified formulation. In doing so, tracklet association can be facilitated by multiple complementary supervision information jointly.

## Unsupervised Visual Representation Learning

Unsupervised visual representation learning aims at learning an effective embedding space from unlabelled data for vision tasks such as image classification (Wu et al. 2018b; Ye et al. 2019; Caron et al. 2018; Huang et al. 2019). For example, Wu et al. (Wu et al. 2018b) propose a non-parametric variation of cross-entropy loss to optimise the model at the instance-level. In (Huang et al. 2019), Huang et al. develop an anchor-based neighbourhood discovery method to improve the robustness of unsupervised feature learning In (Yang, Parikh, and Batra 2016), Yang, Parikh, and Batra introduce a recurrent model to progressively optimise a feature embedding space. In (Ye et al. 2019), Ye et al. jointly exploit data augmentation invariant and instance spread-out property during feature learning. While sharing the same merit of unsupervised visual representation learning, we aim at optimising an effective feature embedding space tailored specially for unsupervised person re-id, a more challenging fine-grained instance recognition problem due to extremely subtle difference between different classes.

# Methodology

## Approach Overview

To fully mine the tracklet structural information for *pure unsupervised* person re-id, we propose a novel tracklet self-supervised learning (TSSL) method. TSSL trains an optimal feature embedding model, enabling both video and image re-id. An overview of TSSL is depicted in Fig. 2.

Given $N$ unlabelled tracklets $\mathcal{X} = \{T_1, T_2, ..., T_N\}$ from any camera views, where each tracklet contains $L$ image frames, *i.e.* $T_i = \{t_{i,j}\}_{j=1}^{L}$, we deploy a feature embedding model $f_\theta(\cdot)$ (where $\theta$ is the network parameters) to extract a feature vector $V_i$ for each tracklet $T_i$ by image-level feature average pooling (*i.e.* $V_{T_i}$), denoted as:

$$V_i = f_\theta(T_i) \tag{1}$$

Without the ground-truth person identity labels, we need to form *self-supervised* learning constraints to facilitate model optimisation. As shown in Fig. 2, we design three types of self-supervision based on the training tracklet data: (1) tracklet frame coherence learning ($\mathcal{L}_f$); (2) tracklet neighbourhood compactness learning ($\mathcal{L}_n$); (3) tracklet cluster structure learning ($\mathcal{L}_c$). These constraints explore the unlabelled training data at different information granularities, yielding strong complementary. The overall model training objective $\mathcal{L}$ is then formulated as:

$$\mathcal{L} = \mathcal{L}_f + \mathcal{L}_n + \mathcal{L}_c \tag{2}$$

At test time, we use the trained $f_\theta(\cdot)$ to extract the feature vector of a *tracklet* or an *image*, and deploy a generic pairwise distance metric $D(\cdot)$ (*e.g.* $L_2$) for re-id matching. We describe the design of the proposed self-supervised learning constraints below.

## Tracklet Frame Coherence Learning

It is intuitive that a frame $t_{i,j}$ from a tracklet $T_i$ should match the source tracklet $T_i$ in the representation space (Chen, Zhu, and Gong 2018), *i.e.* all the frames of a tracklet are coherent. However, directly enforcing this constraint may be sub-optimal. The reason is that, a tracklet $T_i$ is typically short captured in a small time window with limited visual appearance variation across all the constituent frames. To address this problem, for each tracklet we create a positive surrogate with richer appearance variation for generating a stronger frame coherence signal. Specifically, we conduct random image transformation on $T_i$ and generate a positive surrogate tracklet $T_i^* = \{t_{i,j}^*\}_{j=1}^{L}$. Compared with the constituent frames, $T_i^*$ provides more intra-class variation information. One straightforward way is using the mean of all the frames of $T_i^*$ as a positive surrogate. However, this may partly cancel out the transformation effects due to the summation operation. Instead, we randomly select a frame $t_{i,p}^*$ from $T_i^*$ to be a positive surrogate. Also, to enrich variation prediction and to deal with the problem that transformed $t^*_{i,p}$ may be significantly different from $T_i$, which leads to suboptimisation, we reformulate the tracklet as:

$$\mathcal{T}_i = \{\{t_{i,j}\}_{j=1}^{L}, \{t^*_{i,j}\}_{j=1,j\neq p}^{L}\} \tag{3}$$

Based on the triplet loss function (Hermans, Beyer, and Leibe 2017), we formulate the tracklet frame coherence constraint as:

$$\mathcal{L}_f = \max\Big(0, \alpha + D(V_{\mathcal{T}_i}, V_{t^*_{i,p}}) - D(V_{\mathcal{T}_i}, V_{\mathcal{T}_{i,n}})\Big) \tag{4}$$

where $\alpha$ denotes a margin, $\{V_{\mathcal{T}_i}, V_{t^*_{i,p}}, V_{\mathcal{T}_{i,n}}\}$ are the feature vectors of $\{\mathcal{T}_i, t^*_{i,p}, \mathcal{T}_{i,n}\}$ respectively, and $\mathcal{T}_{i,n}$ is a negative tracklet. In design, it is not highly necessary to guarantee 100% accuracy for $\mathcal{T}_{i,n}$, as long as it is statistically reliable so that the model can mine useful information. As the majority is negative, we use the cyclic ranking consistency as (Yang et al. 2019; Chen, Zhu, and Gong 2018; Liu, Wang, and Lu 2017) for quality guarantee.

## Tracklet Neighbourhood Compactness Learning

In (Li, Zhu, and Gong 2018a; Chen, Zhu, and Gong 2018), each tracklet is associated with a camera identity label, which enables cross-view nearest tracklet search. However, this is less scalable because the camera annotation is not always available. To relax this assumption, we form the neighbourhood for each tracklet in the whole training data, without using any camera labels.

From the data manifold perspective, neighbours are likely to share the underlying class label, providing a natural source for self-supervision (Huang et al. 2019). Under this consideration, we formulate a tracklet neighbourhood compactness constraint as:

$$\mathcal{L}_n = -\lambda \log \frac{\exp(-sD(V_{\mathcal{T}_i}, V_{K_i})^2)}{\sum_{j=1,j\neq i}^{N} \exp(-sD(V_{\mathcal{T}_i}, V_{\mathcal{T}_j})^2)} \tag{5}$$

where $s$ specifies a scale parameter, $\lambda$ is a compensation parameter, $K_i$ is a neighbourhood tracklet of $\mathcal{T}_i$ and $V_{K_i}$ is the corresponding feature vector. This encourages the model to pull each tracklet closer to its neighbours. To minimise wrong self-supervision from false neighbours, we only select the nearest neighbour in this constraint. The neighbourhood is established based on pairwise distance between $\mathcal{T}_i$ and all the other tracklets. To computationally facilitate this process, we maintain a global tracklet module $\mathcal{M}$ where the feature vector $V_t$ of a tracklet at the $t$-th iteration is updated as:

$$V_t = (1-\eta)V + \eta V_{t-1} \tag{6}$$

where $V$ denotes the up-to-date feature vector and $\eta$ is the update momentum.

## Tracklet Cluster Structure Learning

Clustering is an effective strategy for unsupervised learning. However, it is non-trivial to apply it for person re-id due to the fine-grained recognition nature with subtle inter-class differences whilst large intra-class variations. To obtain reliable cluster structure for self-supervision, we adopt the agglomerative clustering method as (Lin et al. 2019).

**Algorithm 1** Tracklet Self-Supervised Learning.

**Input:** Unlabelled person tracklet $\mathcal{X}$.
**Output:** A learned feature embedding model $f_\theta(\cdot)$.

```
Initialise: Model parameters $\theta$
for $step = 1 \rightarrow 1/\delta$ do /* Stage level */
    for $e = 1$ →Max-epoch do /* Epoch level */
        for $b = 1$→Batch-number do /* Batch level */
            Construct a positive surrogate tracklet $T^*$
            Forward to get tracklet features $V$ (Eq. (1))
            Compute frame coherence loss $\mathcal{L}_f$ (Eq. (4))
            Update global tracklet module $\mathcal{M}$(Eq. (6))
            Compute neighbourhood loss $\mathcal{L}_n$ (Eq. (5))
            Compute cluster structure loss $\mathcal{L}_c$ (Eq. (7))
            Backward to update $\theta$ with Eq. (2)
            Update the cluster memory $V_c$
        end for
    end for
    Update the clusters with Eq. (8)
    Re-initialise cluster memory $V_c$
    Evaluate the performance of $f_\theta(\cdot)$ as $\mathcal{P}^{step}$
    Track the best model $\theta^* = \theta$ according to $\mathcal{P}^{step}$
end for
return $f_\theta(\cdot) = f_{\theta^*}(\cdot)$
```

Based on a tracklet clustering solution, we then formulate the tracklet cluster constraint as:

$$\mathcal{L}_c = -\log \frac{\exp(V_{c,i}^T V_i/\tau)}{\sum_{j=1}^{N_c} \exp(V_{c,j}^T V_i/\tau)} \tag{7}$$

where $\tau$ is a temperature parameter (Hinton, Vinyals, and Dean 2015), and $N_c$ is the cluster number. $V_c$ is an external memory bank maintaining the feature vectors $V_{c,\cdot}$ for each cluster, which is updated using the scheme in Eq. (6).

In agglomerative clustering, cluster merging at each iteration is a key. To this end, (Lin et al. 2019) uses the minimum pairwise distance between two clusters $V_{c,i}$ and $V_{c,j}$ to represent their distance. They assume that all the clusters are distributed evenly and/or learned in a balanced manner, which is often not true. To address this limitation, we use a new inter-cluster distance measurement that not only further considers the neighbour cluster distribution, but also enables to progressively merge small clusters in an easy-to-hard manner.

Formally, we formulate the proposed distribution-aware cluster pairwise distance for agglomerative clustering as:

$$\begin{aligned}\tilde{\mathbf{D}}(V_{c,i}, V_{c,j}) &= D_c(V_{c,i}, V_{c,j}) + \exp\Big(2D_c(V_{c,i}, V_{c,j}) \\ &- \frac{1}{N_k}(\sum_{l=1}^{N_k} D_c(V_{c,i}, V_{K^i_{c,l}}) + \sum_{l=1}^{N_k} D_c(V_{c,j}, V_{K^j_{c,l}}))\Big)\end{aligned} \tag{8}$$

where the exponential term measures the cluster distribution by modelling the *difference* between the cluster centroid distance of two target clusters ($D_c(V_{c,i}, V_{c,j})$) and the average distance of one target cluster with its $N_k$ neighbour clusters ($\frac{1}{N_k}\sum_{l=1}^{N_k} D_c(V_{c,i}, V_{K^i_{c,l}})$) in the inter-cluster distance space. This in essence accounts for the *density statistics* around the target clusters (Yang, Parikh, and Batra 2016).

Rather than combining all the matched cluster pairs at each iteration, we take a *progressive merging* strategy with the merging rate $\delta \in (0, 1)$ for gradually capturing the complex data cluster structure. This aims to minimise error propagation. The intuition is that, two neighbour clusters in a *sparse* embedding region are more likely to share the same concept, since the current model can already separate them at higher confidence (Huang et al. 2019; Lin et al. 2019). To start this progressive process, we start with treating each individual tracklet as a distinct cluster.

## Summary

Our method is end-to-end trainable. We summarise the training process of TSSL for unsupervised person reidentification in Algorithm 1.

# Experiment

## Datasets and Evaluation Protocol

In the literature, image and video based person re-id benchmarks are usually separately evaluated in unsupervised re-id case, with a few exceptions (Li, Zhu, and Gong 2018a; Lin et al. 2019) that consider the both. This is because all the benchmarks were commonly constructed based on the videos, sharing the same raw sources. In this work, we aim at optimising a feature embedding space for both image and video unsupervised re-id, so we also evaluated both image (Market-1501 (Zheng et al. 2015) and DukeMTMC-ReID (Ristani et al. 2016; Zheng, Zheng, and Yang 2017)) and video (MARS (Zheng et al. 2016) and DukeMTMC-VideoReID (Ristani et al. 2016; Wu et al. 2018a)) datasets. The evaluation statistics are summarised in Table 1 with examples shown in Fig. 3.

Given that we consider the tracklet person re-id setting, during training, for the image datasets we followed (Li, Zhu, and Gong 2018a) where the tracklets are constructed from multi-shot images. For fair comparison with existing alternative methods, we used the standard single-query setting (Zheng et al. 2015; Zheng, Zheng, and Yang 2017) at test time. Unlike (Li, Zhu, and Gong 2018a), we did not annotate camera labels to the tracklets, resulting in a more challenging and scalable unsupervised learning setting.

**Evaluation Metrics.** We used the Cumulative Matching Characteristic (CMC) and mean Average Precision (mAP) as the model performance evaluation metrics.

## Implementation Details

We used ResNet-50 (He et al. 2016) (pre-trained on ImageNet) as the feature embedding network. During training, we set $L = 4/16$ for image/video datasets. To generate the positive surrogate, we used random transformations includes horizontal flip, crop, rotation and colour jittering. Meanwhile, data augmentation by random horizontal flip and crop was also applied on the original training tracklets. Average pooling was used to aggregate the frame-level features into a tracklet representation. We empirically set $\alpha = 2$ for Eq. (4), $\eta = 0.5$ for Eq. (6), $\lambda = 0.1$ and $s = 10$ for Eq. (5), $\tau = 0.1$ for Eq. (7), $\delta = 0.05$. We set $N_k = 4$ for cluster merging. The maximal training epoch was set to 20 for the first step and to 5 for the remaining steps. We used Stochastic Gradient Descent (SGD) as the optimiser with the initial learning rate at 0.01 for the backbone model and a decay of 0.1 after 15 training epochs.

Table 1: The evaluation setting statistics. Market-1501, DukeMTMC-ReID and DukeMTMC-VideoReID are abbreviated as Market, Duke, and DukeVideo, respectively.

| Benchmark | Train ID | Test ID | Image | Tracklet |
|---|---|---|---|---|
| Market | 751 | 750 | 32,668 | - |
| Duke | 702 | 702 | 36,411 | - |
| MARS | 625 | 636 | 1,191,003 | 20,478 |
| DukeVideo | 702 | 702 | 815,420 | 4,832 |

(a) Market (b) Duke (c) MARS (d) DukeVideo

Figure 3: Example person pairs.

## Comparisons with the State-of-the-Art Methods

**Competitors.** We compared our TSSL with 14 state-of-the-art unsupervised re-id methods in three groups: **(1)** six *unsupervised cross-domain* re-id models (TJAIDL (Wang et al. 2018), SPGAN (Deng et al. 2018), PTGAN (Wei et al. 2018), HHL (Zhong et al. 2018), PAUL (Yang et al. 2019), ATNet (Liu et al. 2019a)), **(2)** four *one-shot* re-id models (DGM (Ye et al. 2017), Stepwise (Liu, Wang, and Lu 2017), RACE (Ye, Lan, and Yuen 2018), EUG (Wu et al. 2018a)), and **(3)** four *pure unsupervised* re-id models (TAUDL (Li, Zhu, and Gong 2018a), DAL (Chen, Zhu, and Gong 2018), OIM (Xiao et al. 2017), BUC (Lin et al. 2019)).

**Evaluation on Image Benchmarks.** As shown in Table 2, we have the following observations: **(1)** On Market-1501, our TSSL achieves 43.3% in mAP and 71.2% in rank-1, improving the state-of-the-art performance by 2.1% and 4.5% respectively. Although TAUDL (Li, Zhu, and Gong 2018a) also exploits tracklet association, it merely achieves 41.2% in mAP and 63.7% in rank-1 accuracy, which is clearly inferior to TSSL. **(2)** On DukeMTMC-ReID, our TSSL gets the best rank-1 accuracy (62.2%) and the second best mAP (38.5%). While TAUDL (Li, Zhu, and Gong 2018a) performs better in terms of mAP, it assumes extra camera annotation therefore less scalable than TSSL. On the contrary, TSSL does not use any label annotation whilst still achieves very competitive performance. Besides, compared with BUC (Lin et al. 2019) which mines the global cluster supervision, TSSL significantly improves the performance on both Market and Duke. This demonstrates the effectiveness of our tracklet self-supervised learning idea.

Table 2: Comparisons with the state-of-the-art person re-id methods on Market-1501, DukeMTMC-ReID, Mars and DukeMTMC-VideoReID. The best results are in **bold**. $^\dagger$: Unsupervised cross-domain setting, Market (source) $\Rightarrow$ Duke (target) and Duke (source) $\Rightarrow$ Market (target). $^\star$: Results reported in (Lin et al. 2019).

| Methods | Ref. | **Setting**: Label | Market | | Duke | | MARS | | DukeVideo | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | mAP | R1 | mAP | R1 | mAP | R1 | mAP | R1 |
| TJAIDL$^\dagger$ (Wang et al. 2018) | CVPR18 | **Cross-domain:**<br>Large source domain<br>person ID label | 26.5 | 58.2 | 23.0 | 44.3 | - | - | - | - |
| SPGAN$^\dagger$ (Deng et al. 2018) | CVPR18 | | 26.9 | 58.1 | 26.4 | 46.9 | - | - | - | - |
| PTGAN$^\dagger$ (Wei et al. 2018) | CVPR18 | | - | 38.6 | - | 27.4 | - | - | - | - |
| HHL$^\dagger$ (Zhong et al. 2018) | ECCV18 | | 31.4 | 62.2 | 27.2 | 46.9 | - | - | - | - |
| PAUL$^\dagger$ (Yang et al. 2019) | CVPR19 | | 36.8 | 66.7 | 35.7 | 56.1 | - | - | - | - |
| ATNet$^\dagger$ (Liu et al. 2019a) | CVPR19 | | 25.6 | 55.7 | 24.9 | 45.1 | - | - | - | - |
| DGM (Ye et al. 2017) | ICCV17 | **One-shot:**<br>One-shot ID label<br>per person | - | - | - | - | 16.9 | 36.8 | 33.6 | 42.4 |
| Stepwise (Liu, Wang, and Lu 2017) | ICCV17 | | - | - | - | - | 19.7 | 41.2 | 46.8 | 56.3 |
| RACE (Ye, Lan, and Yuen 2018) | ECCV18 | | - | - | - | - | 22.3 | 41.0 | - | - |
| EUG$^\star$ (Wu et al. 2018a) | CVPR18 | | 22.5 | 49.8 | 24.5 | 45.2 | **42.5** | **62.7** | 63.2 | 72.8 |
| TAUDL (Li, Zhu, and Gong 2018a) | ECCV18 | **Pure unsupervised:**<br>Camera label | 41.2 | 63.7 | **43.5** | 61.7 | 29.1 | 43.8 | - | - |
| DAL (Chen, Zhu, and Gong 2018) | BMVC18 | | - | - | - | - | 23.0 | 49.3 | - | - |
| OIM$^\star$ (Xiao et al. 2017) | CVPR17 | **Pure unsupervised:**<br>No label | 14.0 | 38.0 | 11.3 | 24.5 | 13.5 | 33.7 | 43.8 | 51.1 |
| BUC (Lin et al. 2019) | AAAI19 | | 38.3 | 66.2 | 27.5 | 47.4 | 38.0 | 61.1 | 61.9 | 69.2 |
| **TSSL** | Ours | | **43.3** | **71.2** | 38.5 | **62.2** | 30.5 | 56.3 | **64.6** | **73.9** |

Table 3: Evaluating the self-supervised learning components of TSST on Market-1501. $\mathcal{L}_f$: Tracklet frame coherence learning; $\mathcal{L}_c$: Tracklet cluster structure learning; $\mathcal{L}_n$: Tracklet neighbourhood compactness learning.

| Components | mAP | R1 |
|---|---|---|
| $\mathcal{L}_c$ | 35.1 | 65.8 |
| $\mathcal{L}_f + \mathcal{L}_c$ | 42.5 | 70.7 |
| $\mathcal{L}_n + \mathcal{L}_f + \mathcal{L}_c$ | **43.3** | **71.2** |

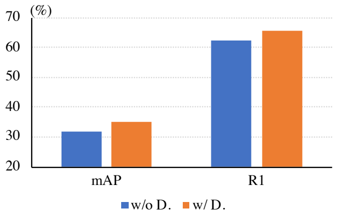

Figure 4: Evaluating the distribution-aware cluster pairwise distance on Market-1501.

**Evaluation on Video Benchmarks.** From Table 2, we observed similar performance comparisons. **(1)** On MARS, TSSL is the second best model in the two pure unsupervised learning groups (inferior to BUC). However, with 30.5% in mAP and 56.3% in rank-1, TSSL outperforms most existing state-of-the-art methods including those one-shot learning models except EUG (Wu et al. 2018a). Compared with TAUDL (Li, Zhu, and Gong 2018a) and DAL (Chen, Zhu, and Gong 2018) which also take the tracklet association idea, TSSL shows better performance. **(2)** On DukeMTMC-VideoReID, TSSL achieves the best mAP (64.6%) and rank-1 accuracy (73.9%), consistently outperforming all unsupervised learning competitors. Overall, these comparisons have comprehensively validated the performance of TSSL.

## Ablation Studies

To further evaluate the proposed TSSL, we conducted a sequence of detailed ablation analyses.

**Self-supervised learning components.** Table 3 shows the impact evaluation of three tracklet self-supervised learning components we proposed in designing TSST. We have several observations: **(1)** With the $\mathcal{L}_c$ alone, our model already achieves fairly strong performance – 35.1% in mAP and 65.8% in rank-1. This verifies the efficacy of the proposed global tracklet cluster structure mining. **(2)** The addition of tracklet frame coherence supervision $\mathcal{L}_f$ significantly improves the performance by 7.4% in mAP and 4.9% in rank-1 accuracy. This demonstrates the significance of our per-tracklet data mining scheme. **(3)** The tracklet neighbourhood compactness constraint $\mathcal{L}_n$ further improves the performance to 43.3% in mAP and 71.2% in rank-1, showing the extra benefit from learning local data structure knowledge. Overall, this test validates the good complementary interaction among the three self-supervised learning signals, leading to strong final model performance collectively.

**Distribution-aware cluster pairwise distance.** To examine the impact of the proposed distribution-aware cluster pairwise distance, we use *only* the tracklet cluster structure learning constraint during model training. As shown in Fig. 4, the proposed distance metric improves the model performance by 3.2% in mAP and 3.3% in rank-1 accuracy, respectively. This validates our design of leveraging the cluster distribution information for more reliable cluster merging in conjunction with the dynamic training process. Furthermore, we incorporate two distance variants (Lin et al. 2019) into TSSL: (1) TSSL+BUC-distance and (2) TSSL+BUC-distance-diversity. As shown in Fig. 5, our distribution-aware TSSL and TSSL+BUC-distance-diversity perform closely, while TSSL+BUC-distance performs worst.

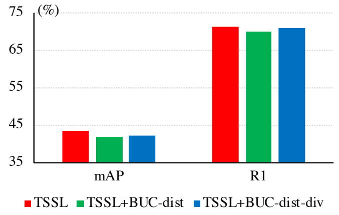

Figure 5: Evaluating distance variants on Market-1501.

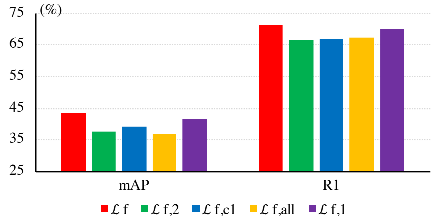

Figure 6: Evaluating different positive surrogate design variants on Market-1501.

**Positive surrogate design.** Apart from the proposed design for positive surrogate used in the tracklet frame coherence learning, we further explored and compared four other formulations: **(1)** two stream surrogate $\mathcal{L}_{f,2}$, *i.e.* formulate a surrogate tracklet as another input stream in a Siamese architecture, **(2)** one constituent frame surrogate $\mathcal{L}_{f,c1}$, *i.e.* use a constituent frame as a surrogate, **(3)** all frame surrogate $\mathcal{L}_{f,all}$, *i.e.* aggregate all the frame features of a surrogate tracklet, and **(4)** single instance surrogate $\mathcal{L}_{f,1}$, *i.e.* randomly select one frame from the tracklet. As shown in Fig. 6, the proposed surrogate design $\mathcal{L}_f$, which exploits random transformation for increasing the intra-class variation, performs the best compared with all the other variants in both mAP and rank-1 accuracy.

**Compensation parameter $\lambda$.** In tracklet neighbourhood compactness learning, we employ a scale parameter in Eq. (8) as (Wang et al. 2017; Yang et al. 2019), which results that $\mathcal{L}_n$ is obviously larger than $\mathcal{L}_c$ and $\mathcal{L}_f$, so we set $\lambda = 0.1$ to make a balance among all the tracklet self-supervision loss components. To test its effect, we varied $\lambda$ in the range from 0 to 0.3 and evaluated their model performance. As shown in Fig. 7, the performance of TSSL increases to 71.2% in rank-1 accuracy and 43.3% in mAP from $\lambda = 0$ to $\lambda = 0.1$, and gradually decreases from $\lambda = 0.1$ to $\lambda = 0.3$. This implies that local neighbourhood constraint needs to avoid over-confidence.

**Neighbour cluster size.** For the neighbour cluster size in the inter-cluster distance space, we take a conservative strategy that only uses $N_k = 4$. We also tested varying sized neighbourhoods. Fig. 8 shows that the rank-1 performance are close when neighbourhoods are sized between 1 and 4, and using more neighbour tracklets tend to decrease the performance due to more false matches are likely to be included. Overall, this verifies that our neighbourhood size selection is both effective and efficient with a low risk of error inclusion.

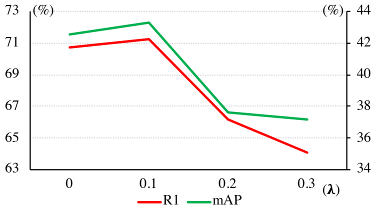

Figure 7: Evaluating the compensation parameter of the tracklet neighbourhood learning on Market-1501.

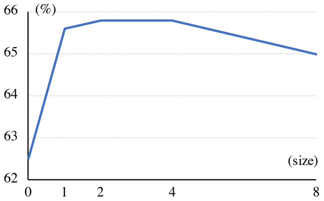

Figure 8: Evaluating the neighbour cluster size on Market-1501.

## Conclusion

In this work, we presented a Tracklet Self-Supervised Learning (TSSL) method for unsupervised image and video person re-id. It enables effective model learning by fully exploiting the underlying discriminative structural information of abundant unlabelled tracklet training data by automatic self-supervision mining at distinct knowledge granularities. Doing so allows to maximise the scalability and usability of TSSL in arbitrarily unconstrained domains. This eliminates not only the expensive cross-camera person identity annotation as required by conventional supervised learning methods, but also the source domain supervision as required by unsupervised cross-domain adaptation methods, and the camera view prior knowledge as required by existing unsupervised tracklet association methods. Extensive experiments on both image and video re-id benchmarks show the superiority of the proposed model against related state-of-the-art methods. We also provided in-depth ablation analyses to examine the impact and efficacy of the proposed component designs in the TSSL formulation.

**Acknowledgements.** This work is supported by Queen Mary University of London Principal's Scholarship, Vision Semantics Limited, Alan Turing Institute Turing Fellowship, and Innovate UK Industrial Challenge Project on Developing and Commercialising Intelligent Video Analytics Solutions for Public Safety (98111-571149).

## References

Caron, M.; Bojanowski, P.; Joulin, A.; and Douze, M. 2018. Deep clustering for unsupervised learning of visual features. In *ECCV*.

Chen, Y.; Zhu, X.; and Gong, S. 2018. Deep association learning for unsupervised video person re-identification. In *BMVC*.

Deng, W.; Zheng, L.; Ye, Q.; Kang, G.; Yang, Y.; and Jiao, J. 2018. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *CVPR*.

Dong, Q.; Gong, S.; and Zhu, X. 2019. Person search by text attribute query as zero-shot learning. In *ICCV*.

Fu, Y.; Wei, Y.; Zhou, Y.; Shi, H.; Huang, G.; Wang, X.; Yao, Z.; and Huang, T. 2019. Horizontal pyramid matching for person re-identification. In *AAAI*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

Hermans, A.; Beyer, L.; and Leibe, B. 2017. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Huang, J.; Dong, Q.; Gong, S.; and Zhu, X. 2019. Unsupervised deep learning by neighbourhood discovery. In *ICML*.

Li, M.; Zhu, X.; and Gong, S. 2018a. Unsupervised person re-identification by deep learning tracklet association. In *ECCV*.

Li, W.; Zhu, X.; and Gong, S. 2018b. Harmonious attention network for person re-identification. In *CVPR*.

Liao, S.; Hu, Y.; Zhu, X.; and Li, S. Z. 2015. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*.

Lin, Y.; Dong, X.; Zheng, L.; Yan, Y.; and Yang, Y. 2019. A bottom-up clustering approach to unsupervised person re-identification. In *AAAI*.

Liu, J.; Zha, Z.-J.; Chen, D.; Hong, R.; and Wang, M. 2019a. Adaptive transfer network for cross-domain person re-identification. In *CVPR*.

Liu, Y.; Yuan, Z.; Zhou, W.; and Li, H. 2019b. Spatial and temporal mutual promotion for video-based person re-identification. In *AAAI*.

Liu, Z.; Wang, D.; and Lu, H. 2017. Stepwise metric promotion for unsupervised video person re-identification. In *ICCV*.

Ristani, E.; Solera, F.; Zou, R.; Cucchiara, R.; and Tomasi, C. 2016. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV*.

Tesfaye, Y. T.; Zemene, E.; Prati, A.; Pelillo, M.; and Shah, M. 2019. Multi-target tracking in multiple non-overlapping cameras using fast-constrained dominant sets. *IJCV*.

Wang, F.; Xiang, X.; Cheng, J.; and Yuille, A. L. 2017. Normface: l 2 hypersphere embedding for face verification. In *ACMMM*.

Wang, J.; Zhu, X.; Gong, S.; and Li, W. 2018. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *CVPR*.

Wei, L.; Zhang, S.; Gao, W.; and Tian, Q. 2018. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*.

Wu, Y.; Lin, Y.; Dong, X.; Yan, Y.; Ouyang, W.; and Yang, Y. 2018a. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In *CVPR*.

Wu, Z.; Xiong, Y.; Yu, S. X.; and Lin, D. 2018b. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*.

Wu, G.; Zhu, X.; and Gong, S. 2019a. Person re-identification by ranking ensemble representations. In *ICIP*.

Wu, G.; Zhu, X.; and Gong, S. 2019b. Spatio-temporal associative representation for video person re-identification. In *BMVC*.

Xiao, T.; Li, S.; Wang, B.; Lin, L.; and Wang, X. 2017. Joint detection and identification feature learning for person search. In *CVPR*.

Yang, Q.; Yu, H.-X.; Wu, A.; and Zheng, W.-S. 2019. Patch-based discriminative feature learning for unsupervised person re-identification. In *CVPR*.

Yang, J.; Parikh, D.; and Batra, D. 2016. Joint unsupervised learning of deep representations and image clusters. In *CVPR*.

Ye, M.; Ma, A. J.; Zheng, L.; Li, J.; and Yuen, P. C. 2017. Dynamic label graph matching for unsupervised video re-identification. In *ICCV*.

Ye, M.; Zhang, X.; Yuen, P. C.; and Chang, S.-F. 2019. Unsupervised embedding learning via invariant and spreading instance feature. In *CVPR*.

Ye, M.; Lan, X.; and Yuen, P. C. 2018. Robust anchor embedding for unsupervised video person re-identification in the wild. In *ECCV*.

Yu, H.-X.; Zheng, W.-S.; Wu, A.; Guo, X.; Gong, S.; and Lai, J.-H. 2019. Unsupervised person re-identification by soft multilabel learning. In *CVPR*.

Zhang, X.; Cao, J.; Shen, C.; and You, M. 2019a. Self-training with progressive augmentation for unsupervised cross-domain person re-identification. In *ICCV*.

Zhang, Y.; Zhong, Q.; Ma, L.; Xie, D.; and Pu, S. 2019b. Learning incremental triplet margin for person re-identification. In *AAAI*.

Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; and Tian, Q. 2015. Scalable person re-identification: A benchmark. In *ICCV*.

Zheng, L.; Bie, Z.; Sun, Y.; Wang, J.; Su, C.; Wang, S.; and Tian, Q. 2016. MARS: A video benchmark for large-scale person re-identification. In *ECCV*.

Zheng, Z.; Zheng, L.; and Yang, Y. 2017. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*.

Zhong, Z.; Zheng, L.; Li, S.; and Yang, Y. 2018. Generalizing a person retrieval model hetero-and homogeneously. In *ECCV*.

Zhong, Z.; Zheng, L.; Luo, Z.; Li, S.; and Yang, Y. 2019. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *CVPR*.