Exploiting Temporal Coherence for Self-Supervised One-shot Video Re-identification

Dripta S. Raychaudhuri and Amit K. Roy-Chowdhury

University of California, Riverside
{draychaudhuri,amitrc}@ece.ucr.edu

Abstract. While supervised techniques in re-identification are extremely effective, the need for large amounts of annotations makes them impractical for large camera networks. One-shot re-identification, which uses a singular labeled tracklet for each identity along with a pool of unlabeled tracklets, is a potential candidate towards reducing this labeling effort. Current one-shot re-identification methods function by modeling the inter-relationships amongst the labeled and the unlabeled data, but fail to fully exploit such relationships that exist within the pool of unlabeled data itself. In this paper, we propose a new framework named Temporal Consistency Progressive Learning, which uses temporal coherence as a novel self-supervised auxiliary task in the one-shot learning paradigm to capture such relationships amongst the unlabeled tracklets. Optimizing two new losses, which enforce consistency on a local and global scale, our framework can learn learn richer and more discriminative representations. Extensive experiments on two challenging video re-identification datasets - MARS and DukeMTMC-VideoReID - demonstrate that our proposed method is able to estimate the true labels of the unlabeled data more accurately by up to 8%, and obtain significantly better re-identification performance compared to the existing state-of-the-art techniques.

Keywords: video person re-identification, temporal consistency, one-shot learning, semi-supervised learning

1 Introduction

Person re-identification (re-ID) aims to solve the challenging problem of matching identities across non-overlapping views in a multi-camera system. The surge of deep neural networks in computer vision [13, 25] has been reflected in person re-ID as well, with impressive performance over a wide variety of datasets [28, 5]. However, this improved performance has predominantly been achieved through *supervised learning*, facilitated by the availability of large amounts of annotated data. However, acquiring identity labels for a large set of unlabeled tracklets is an extremely time-consuming and cumbersome task. Consequently, methods which can ameliorate this annotation problem and work with limited supervision, such as *unsupervised learning* or *semi-supervised learning* techniques, are of primary importance in the context of person re-ID.

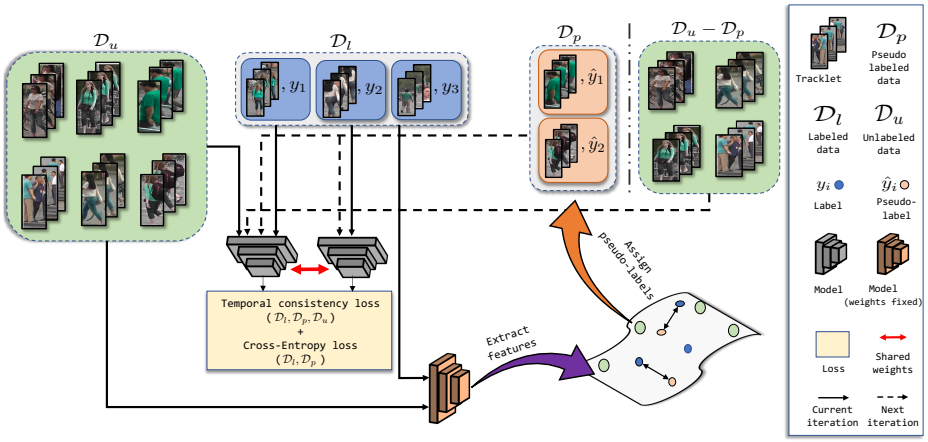


Fig. 1. A schematic illustration of the proposed framework. Our method makes use of both labeled and unlabeled tracklets at every iteration of model training. The first step involves learning the parameters of the deep model by using temporal consistency as self-supervision and, additionally, softmax loss on the minimal set of annotated tracklets. Next, this model is used to predict pseudo-labels on a few confident samples. These two steps alternate, one after the other, until the entire unlabeled set has been incorporated in terms of pseudo-labels.

In this work, we focus on the semi-supervised task in video person re-ID, specifically, the *one-shot* setting, where only one tracklet per identity is labeled. The objective of the learning process is to utilize this small labeled set along with a larger unlabeled set of tracklets to obtain a re-ID model. The key challenge involved with the one-shot task is figuring out the inter-relationships which exist amongst the labeled and unlabeled instances. State-of-the-art one-shot methods try to address this by estimating the labels of the unlabeled tracklets (pseudo-labels) and then utilizing a supervised learning strategy. Some works employ a static sampling strategy [32, 18], where pseudo-labels with a confidence score above a pre-defined threshold are selected for supervised learning. More recent works [30, 29] make use of a progressive sampling strategy, where a subset of the pseudo-labeled samples are selected with the size of the subset expanding with each iteration. This prevents an influx of noisy pseudo-labels, and thus, averts the situation of confirmation bias [1]. However, in an effort to control the number of noisy pseudo-labels, most of these methods discard a significant portion of the unlabeled set at each learning iteration; thus, the information in the unlabeled set is not maximally utilized for training the model. Due to this inefficient usage of the unlabeled set and the limited number of labeled instances, propagating beliefs directly from the labeled to the unlabeled set is insufficient to fully capture the relationships which exist amongst instances of the unlabeled set.

To resolve this issue of inefficient usage of the unlabeled data, we draw inspiration from the field of self-supervised visual representation learning [12].

We propose using *temporal coherence* [24, 21, 19] as a form of self-supervision to maximally utilize the unlabeled data and learn discriminative person specific representations. Temporal coherence is motivated by the fact that features corresponding to a person in a tracklet should be focused on the discriminative aspects related to the person, such as clothing and gait, and ignore background nuances such as illumination and occlusion (see Fig. 2). This naturally suggests that features should be temporally consistent across the entire duration of the tracklet as the person in a tracklet remains constant. Thus, we propose a new framework, *Temporally Consistent Progressive Learning* (TCPL), which unifies this notion of temporal coherence with a progressive pseudo-labeling strategy [30]. An overview of our framework is presented in Fig. 1.

In this paper, we propose two novel losses to learn such temporally consistent features: *Intra-sequence temporal consistency loss* and the *Inter-sequence temporal consistency loss*. Both of these losses apply consistency regularization on the temporal dimension of a tracklet. While the first loss employs a local level of consistency by operating on a specific tracklet, the second loss extends it by applying temporal consistency both *within and across* tracklets.

Using such self-supervised losses, our framework can use the unlabeled data at each iteration of learning, allowing maximal information to be extracted out of it. Additionally, by exploiting two levels of consistency, as explained above, TCPL can better model the relationships amongst the unlabeled instances without being limited by the number of labeled instances. Thus, our framework addresses both the drawbacks associated with the current crop of methods and achieves state-of-the-art performance in the one-shot person re-ID task.

Main contributions. Our main contributions are summarised as follows:

- We introduce a new framework, *Temporally Consistent Progressive Learning*, which unifies self-supervision and pseudo-labeling to maximally utilize the labeled and unlabeled data efficiently for one-shot video person re-ID.
- We introduce two novel self-supervised losses, the *Intra-sequence temporal consistency loss* and the *Inter-sequence temporal consistency loss*, to implement temporal consistency and empirically demonstrate their benefits in learning richer and more discriminative feature representations.
- We demonstrate that this intelligent use of the unlabeled data through self-supervision, unlike previous pseudo-labeling methods, leads to significantly better label estimation and superior results on the one-shot video re-ID task, outperforming the state-of-the art one-shot video re-ID methods on the MARS and DukeMTMC-VideoReID datasets.

2 Related works

The majority of the literature in person re-ID has focused on *supervised* learning on labeled images/tracklets of persons [38, 37, 4, 31]. While these techniques achieve excellent results on many datasets, they require a substantial amount of annotations. The need to alleviate this excessive need for labeled data has led to

research into *unsupervised* [34, 35, 16, 6] and *semi-supervised* [30, 29, 7] methods. We provide a review of the relevant developments in these fields. In addition, our work draws inspiration from the ideas explored in the domain of *self-supervision*.

Unsupervised person re-ID. Recent unsupervised methods [34, 35, 16, 6] mostly use some form of deep clustering. The authors in [15] utilise a camera aware loss by defining nearest neighbors across cameras as being similar. In [16], an agglomerative clustering scheme is introduced, alternating between learning of features and clustering using the learnt features. However, these methods still lag behind supervised methods by quite some distance. Another line of research utilises auxiliary datasets, which are completely labeled, for initializing a re-ID model and then using unsupervised domain adaptation techniques on the unsupervised target dataset.

Semi-supervised & one-shot person re-ID. The unsatisfactory performance of purely unsupervised methods [34, 35, 16, 6] has given rise to semi-supervised and one-shot methods in re-ID. Some of the major ideas utilized in these methods include dictionary learning [17], graph matching [10] and metric learning [2]. More recently, new methods in this setting try to estimate the labels of the unlabeled tracklets (pseudo-labels) with respect to the labeled tracklets and then utilise a supervised learning strategy. The authors of [32] use a dynamic graph matching strategy which iteratively updates the image graph and the label estimation to learn a better feature space with intermediate estimated labels. A stepwise metric learning approach to the problem is proposed in [18]. Both these methods employ a static sampling strategy, where pseudo-labels with a confidence score above a pre-defined threshold are selected at each step - this leads to a lot of noisy labels being incorporated and hinders the learning process due to due to *confirmation bias* [1]. In order to contain the noise, the authors of [30, 29] approach the problem from a progressive pseudo-label selection strategy, where the subset of the pseudo-labeled samples selected gradually increase with iterations. While this prevented the influx of noisy pseudo-labels, a significant portion of the unlabeled set is discarded at each step and thus, the unlabeled set is used inefficiently. We address this issue by using self-supervision.

Self-supervised learning. Self-supervised learning utilizes pretext tasks, formulated using only unsupervised data. A pretext task is designed in a such a way that solving it requires the model to learn useful visual features. These tasks can involve predicting the angle of rotation applied to an image [9] or predicting a permutation of multiple randomly sampled and permuted patches [22]. Some techniques go beyond solving such auxiliary classification tasks and enforce constraints on the representation space. A prominent example is the exemplar loss from [8]. Our method belongs to this latter category of self-supervision and imposes temporal consistency on tracklet features.

3 Methodology

In this section, we present our framework (TCPL) for solving the task of one-shot video person re-ID. First, we provide a background on the current progressive

pseudo-labeling methods and discuss their shortcomings. Thereafter, we turn to our proposed temporal consistency losses and describe their workings, before presenting our integrated framework. Before going into the details of our framework, let us define the notations and problem statement formally.

Problem statement. Consider that we have a training set of m tracklets, $\mathcal{D} = \{\mathcal{X}_i\}_{i=1}^{m}$, which are acquired from a camera network. One-shot re-ID assumes that there exists a set $\mathcal{D}_l \subset \mathcal{D}$, which contains a singular labeled tracklet for each identity. Thus, $\mathcal{D}_l = \{(\mathcal{X}_i, y_i)\}_{i=1}^{m_l}$, where $y_i \in \{0,1\}^{m_l}$ such that y_i is 1 only at dimension i and 0 otherwise, and m_l denotes the number of distinct identities. The rest of the tracklets, $\mathcal{D}_u = \mathcal{D} - \mathcal{D}_l = \{\mathcal{X}_i\}_{i=1}^{m_u}$ do not possess annotations. Our goal is to learn a discriminative person re-ID model $f_\theta(\cdot)$ utilizing both $\mathcal{D}_l$ and $\mathcal{D}_u$. During inference, $f_\theta(\cdot)$ is used to embed both the probe $\mathcal{X}^{\mathtt{q}}$ and gallery tracklets $\{\mathcal{X}_i^{\mathtt{g}}\}_{i=1}^{m_g}$ into a common space and then rank all the gallery tracklets by evaluating their degree of correspondence to the probe via some metric. What makes this challenging, even more so than the semi-supervised task, is the fact that $m_l \ll m_u$ and each identity has only a single labeled tracklet.

3.1 Progressive Pseudo-labeling and its drawbacks

The progressive pseudo-labeling paradigm is an enhancement over the original pseudo-labeling framework [14] where one imputes approximate classes on unlabeled data by making predictions from a model trained only on labeled data. The learning process involves the following two steps for each step of learning: (1) train the model via supervised learning on the labeled data and the pseudo-labeled data; (2) select a few reliable pseudo-labeled candidates from unlabeled data according to a prediction reliability criterion.

In [30], the authors gradually select larger sets of pseudo-labeled data to be incorporated into the supervised learning process via a dissimilarity criterion. Pseudo labels are assigned to the unlabeled candidates by the identity labels of their nearest labeled neighbors in the embedding space. The distance to the corresponding labeled neighbor is designated as the dissimilarity cost, which is used as the measure of reliability for the pseudo label. However, as a result of the strict selection criterion, this does not use the unlabeled set efficiently - discarding a significant amount of unlabeled data at each step of pseudo labeling.

To improve the efficiency, the authors in [29] propose to set up a memory bank to store the instance features $v_i = f_\theta(\mathcal{X}_i)$ calculated in the previous step. Then the probability of sample $\mathcal{X}_j$ being recognized as the i-th instance can be written as,

$$P(i|\mathcal{X}_j) = \frac{\exp\left(v_i^T f_\theta(\mathcal{X}_j)/\tau\right)}{\sum_k \exp\left(v_k^T f_\theta(\mathcal{X}_j)/\tau\right)} \tag{1}$$

where τ is the temperature parameter controlling the softness of the distribution. Minimizing the negative log-likelihood of $\sum_i P(i|\mathcal{X}_j)$, which they call the *exclusive loss*, pulls each instance $\mathcal{X}_i$ towards its corresponding memorized vector v_i and repels the memorized vectors of other instances. Due to efficiency issues, the

memorized feature v_i corresponding to instance $\mathcal{X}_i$ is only updated in the iteration which takes $\mathcal{X}_i$ as input [33]. In other words, the memorized feature v_i is only updated once per epoch. However, the network itself is updated in each iteration, rendering the memory bank scheme inefficient. In addition, the exclusive loss looks at the global data distribution, similar to the softmax loss, forcing embeddings corresponding to different identities to stay apart for encouraging inter-class separability. The local data distribution or the intra-class similarity, is left unaddressed and thus, the improvement over softmax is negligible.

In the next section, we present how temporal coherence can be employed to amend these drawbacks.

3.2 Temporal coherence as self-supervision

In the previous section, we discussed the two fundamental problems plaguing the current crop of progressive pseudo-labeling methods: (1) inefficient usage of the unlabeled set, (2) focusing strictly on the global data distribution. To ameliorate these drawbacks, *we propose to use temporal coherence as a form of self-supervision.* Consistency across the frames in a tracklet encourages the model to focus on the *local* distribution of the data and learn features which incorporate the specific attributes of the individual in the tracklet and ignore spurious artifacts such as background and lighting variation. This also provides a straightforward approach towards utilizing the entire unlabeled set, irrespective of whether some specific unlabeled instance is assigned a confident pseudo-label. In the following sections, we present two novel losses: *Intra-sequence temporal consistency* and *Inter-sequence temporal consistency*, which implement this notion of temporal consistency and show how to integrate them into a self-learning framework towards solving the one-shot video re-ID task.

Intra-sequence temporal consistency. The intra-sequence temporal consistency loss is based on the idea of video temporal coherence [24, 21, 19]. While the previous works focus on learning the temporal order by considering individual frames, we use consistency as a tool for the learnt features to implicitly *ignore background nuances* and *focus on the actual person attributes.* We do this by sampling non-overlapping mini-tracklets from a tracklet and enforce the embeddings corresponding to these mini-tracklets to come closer via a contrastive loss.

Given a tracklet $\mathcal{X}$ consisting of frames $\{x_1, \cdots, x_n\}$, intra-sequence consistency involves creating two mini-tracklets $\mathcal{X}^{\mathsf{a}}$ and $\mathcal{X}^{\mathsf{p}}$ by sampling two mutually exclusive sets of frames from the original tracklet $\mathcal{X}$. This is done by the function $\Phi_{\mathsf{T}}(\mathcal{X})$, which first divides the $\mathcal{X}$ into a set of mini-tracklets, each of size $\rho \cdot |\mathcal{X}|$ and then samples from it as follows,

$$\mathcal{X}^{\mathsf{a}}, \mathcal{X}^{\mathsf{p}} = \Phi_{\mathsf{T}}(\mathcal{X}) \tag{2}$$

More specifically, $\Phi_{\mathsf{T}}(\mathcal{X})$ samples from the set $\{\mathcal{X}^1, \mathcal{X}^2, \cdots, \mathcal{X}^{1/\rho}\}$ uniformly without replacement. Here, ρ is a hyper-parameter that controls the size of each mini-tracklet with respect to the size of the tracklet $|\mathcal{X}|$. This ensures that

$\mathcal{X}^{\mathtt{a}} \cap \mathcal{X}^{\mathtt{p}} = \emptyset$, and consequently, these tracklets are temporally incoherent. For all our experiments, ρ is set to 0.2. After obtaining these tracklets the loss forces their respective representations to be consistent temporally with one another as follows,

$$\mathcal{L}_{\text{intra}} = \|f_\theta(\mathcal{X}^{\mathtt{a}}) - f_\theta(\mathcal{X}^{\mathtt{p}}))\|_2. \tag{3}$$

This definition of the intra-sequence temporal consistency can be interpreted as a from of consistency regularization [20, 27, 23], which measures discrepancy between predictions made on perturbed unlabeled data points, i.e.,

$$\mathcal{L}_{\text{cons}} = d\left(p(y|x), p(y|\hat{x})\right) \tag{4}$$

where $d(\cdot, \cdot)$ is a divergence measure and $\hat{x} = x + \delta$. Such regularization focuses on the local data distribution, and implicitly pushes the decision boundary away from high-density parts of the unlabeled data to enhance intra-class similarity in accordance to the *cluster assumption* [3]. In our formulation, the two mini-tracklets are *temporally perturbed versions of each other in terms of background*, i.e., $x = \mathcal{X}^{\mathtt{a}}$, $\hat{x} = \mathcal{X}^{\mathtt{p}}$ and δ indicates perturbations in time - the consistency is applied on features, instead of distributions, and across time.

Inter-sequence temporal consistency. The intra-sequence temporal consistency loss focuses solely on the intra-class similarity. To learn a discriminative person re-ID model, the learning process also has to account for the global distribution of the data or the inter-class separability. The triplet loss [11] has been widely used in the re-identification and retrieval literature for its ability to encode such global information.

The triplet loss ensures that, given an anchor point $\mathcal{X}^{\mathtt{a}}$, the feature of a positive point $\mathcal{X}^{\mathtt{p}}$ belonging to the same class (person) y_a is closer to the feature of the anchor than that of a negative point $\mathcal{X}^{\mathtt{n}}$ belonging to another class y_n, by at least a margin α. However, directly using the triplet loss is not possible in our scenario as it uses identity label information and thus, its effectiveness will depend heavily on the quality of label estimation. Therefore, we propose the inter-sequence temporal consistency loss, which induces a global level of consistency similar to the standard triplet formulation *without access to labels*.

Specifically, given a tracklet $\mathcal{X}$, we sample two temporally incoherent mini-tracklets in the same manner as mentioned in the previous section. Without loss of generality, we treat one as the anchor $\mathcal{X}^{\mathtt{a}}$, and the other one as the positive point $\mathcal{X}^{\mathtt{p}}$, which contains the same identity, but temporally perturbed. For the negative instance, we obtain it from the batch nearest neighbors of $\mathcal{X}^{\mathtt{a}}$. This is done by creating the corresponding ranking list of tracklets in the batch B, excluding $\mathcal{X}$ and sampling a tracklet $\mathcal{X}^{\mathtt{n}}$ uniformly within the range of ranks $[r, 2r]$ as follows:

$$\mathcal{X}^{\mathtt{n}} = \Psi(\mathcal{N}_{[r,2r]}(\mathcal{X})) \tag{5}$$

where $\Psi(\cdot)$ denotes sampling from a set of elements uniformly. $\mathcal{N}_{[r,2r]}(\mathcal{X})$ indicates the nearest neighbors of $\mathcal{X}$ in the batch (up to a total of B neighbors) which are

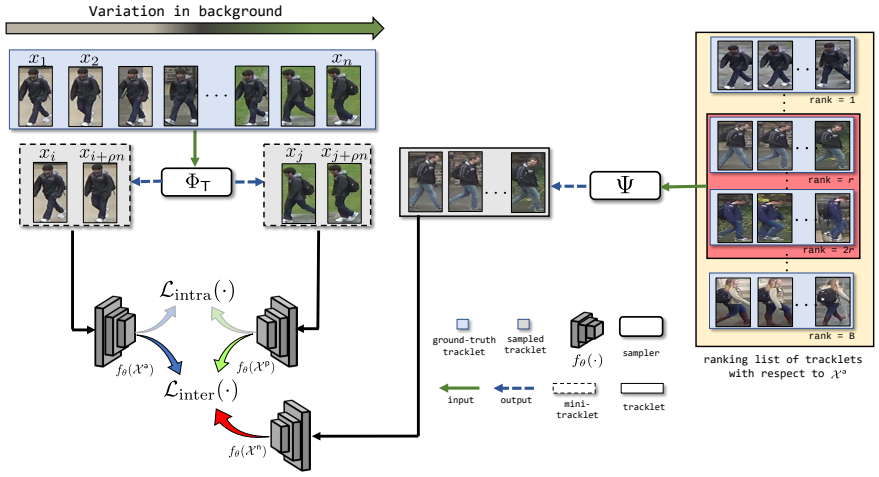


Fig. 2. An illustration of the inter-sequence temporal consistency criterion. Firstly, we sample temporally incoherent mini-tracklets using Φ_T to serve as the anchor and positive sample. Note the temporal perturbations in these mini-tracklets, manifested in the form changing background. Next, Ψ is used to obtain the negative sample from the batch nearest neighbors of the anchor, using a ranking based criterion. Using these, we formulate the triplet loss to enforce consistency such that $f_\theta(\cdot)$ learns features which focus on the discriminative aspects related to the person in the tracklet and ignore the background nuances.

ranked in the range $[r, 2r]$. Using this range of ranks we filter out the possible positive samples and the easy negative samples, which are very low in the ranking list and potentially contribute to zero gradient. This strategy allows us to choose potential hard negatives which have been shown to give best performance [11]. The value of r is set to 3 and α to 0.3, for all our experiments.

Thus, the inter-sequence temporal consistency loss can be formulated as,

$$\mathcal{L}_{\text{inter}} = \max\{0, \|f_\theta(\mathcal{X}^{\mathtt{a}}) - f_\theta(\mathcal{X}^{\mathtt{p}})\|_2 - \|f_\theta(\mathcal{X}^{\mathtt{a}}) - f_\theta(\mathcal{X}^{\mathtt{n}})\|_2 + \alpha\} \tag{6}$$

A pictorial representation of the loss formulation is presented in Fig. 2.

3.3 Temporal Consistency Progressive Learning

In this section, we present our proposed framework, *Temporal Consistency Progressive Learning* (TCPL), based on the temporal consistency self-supervised losses discussed in the previous section. TCPL integrates self-supervision with pseudo-labeling to learn the person re-ID model. Temporal coherence is used to enhance the feature learning process in the form of multi-task learning. Training of this framework alternates between two key steps: (1) Representation learning, (2) Assignment of pseudo-labels.

Representation learning. In order to learn the weights of the embedding function $f_\theta(\cdot)$, we jointly optimize the following loss function,

$$\mathcal{L} = \sum_{(\mathcal{X},y)\in\mathcal{D}_l} \mathcal{L}_l(\mathcal{X},y) + \sum_{(\mathcal{X},\hat{y})\in\mathcal{D}_p} \mathcal{L}_l(\mathcal{X},\hat{y}) + \lambda \left(\sum_{\mathcal{X}\in\mathcal{D}} \mathcal{L}_{\text{intra}}(\mathcal{X}) + \sum_{\mathcal{X}\in\mathcal{D}} \mathcal{L}_{\text{inter}}(\mathcal{X}) \right) \quad (7)$$

where $\mathcal{L}_l$ is a standard cross-entropy classification loss applied on all labeled and selected pseudo-labeled tracklets in the dataset. The supervised loss $\mathcal{L}_l$ is optimized by appending a classifier $g_W(\cdot)$ on top of the feature extractor $f_\theta(\cdot)$ as

$$\mathcal{Z} = g_W(f_\theta(\mathcal{X})) = W^T f_\theta(\mathcal{X}) + b \quad (8)$$

$$\mathcal{L}_l = -\log\left(\frac{e^{y^T\mathcal{Z}}}{\sum_j e^{\mathcal{Z}_j}}\right), \quad (9)$$

where $f_\theta(\mathcal{X}) \in \mathbb{R}^{d\times 1}$, $W \in \mathbb{R}^{d\times m_l}$ and $b \in \mathbb{R}^{m_l\times 1}$. The value of d represents the feature dimension and is equal to 2048 in our experiments. The labeled set and pseudo-labeled set are denoted by $\mathcal{D}_l$ and $\mathcal{D}_p$ respectively, with $\hat{y}$ denoting the pseudo-labels, while $\mathcal{D}$ refers to the entire set of tracklets. Note that, $\mathcal{D}_l \subset \mathcal{D}$ and $\mathcal{D}_p \subset \mathcal{D}$, such that $\mathcal{D}_p \cap \mathcal{D}_l = \emptyset$. The hyper-parameter λ is a non-negative scalar that controls the weight of temporal consistency in the joint loss function.

Assignment of pseudo-labels. Following [30], we use the nearest neighbor in the embedding space to assign pseudo-labels - each unlabeled tracklet is assigned a pseudo-label by transferring the label of its nearest labeled neighbor in the embedding space. For $\mathcal{X}_j \in \mathcal{D}_u$,

$$i = \arg\min_{\mathcal{X}_k\in\mathcal{D}_l} \|f_\theta(\mathcal{X}_j) - f_\theta(\mathcal{X}_k)\|_2, \quad (10)$$

$$\hat{y}_j = y_i \quad (11)$$

After assignment of the pseudo-labels, a confidence criterion is used to choose the most reliable predictions to be used in optimizing $\mathcal{L}_l$ for the next step. Instead of a static threshold, a total of n_t samples are selected at step t by choosing the top n_t unlabeled samples with smallest distance to their corresponding labeled nearest neighbour and added to $\mathcal{D}_p$. A smaller value of the distance implies a more confident pseudo-label prediction.

The value of n_t is incremented gradually with t, depending on an enlarging factor $p \in (0,1)$ [30] where, $n_t = n_{t-1} + pn_u$. Thus, the learning process continues for a total of $(\lfloor 1/p \rfloor + 1)$ steps - until the entire unlabeled set has been assigned confident pseudo-labels. The parameter p controls the trade-off between label estimation accuracy and training time - a smaller value of p leads to better label estimation at the cost of higher training time.

4 Experiments

We evaluate our proposed method on two popular video person re-ID benchmarks, namely, MARS [36] and DukeMTMC-VideoReID [26]. MARS is the largest video

Algorithm 1 Temporally Consistent Progressive Learning

INPUT: Labeled set $\mathcal{D}_l$, unlabeled set $\mathcal{D}_u$, enlarging ratio p, sampling factor ρ, loss weight λ, randomly initialized model $f_{\theta_0}(\cdot)$
OUTPUT: Feature extractor $f_{\theta_{opt}}(\cdot)$

Initialize the selected pseudo-labeled data $\mathcal{D}_p^0 \leftarrow \emptyset$, step $t \leftarrow 0$, sampling size $n_0 \leftarrow 0$, $n_u = |\mathcal{D}_u|$
while $n_t \leq n_u$ **do**
  $t \leftarrow t + 1$
  Train the model using (7)
  Assign pseudo-labels using (10)
  $n_t \leftarrow n_{t-1} + p.n_u$
  Choose the n_t most confident pseudo-labels and add to $\mathcal{D}_p^{t-1}$
end while
Choose model with best validation performance

re-ID dataset containing 17,503 tracklets for 1,261 identities and 3,248 distractor tracklets, which are captured by six cameras. The DukeMTMC-VideoReID dataset is captured using 8 cameras and contains 2,196 tracklets for training and 2,636 tracklets for testing. Standard splits are used along with distractors.

Evaluation metrics. Given a probe tracklet, we calculate the Euclidean distance with respect to all the gallery tracklets, and sort the distances to obtain the final ranking list. We utilize the Cumulative Matching Characteristics (CMC) and mean Average Precision (mAP) as the performance evaluation measures. We report the Rank-1, Rank-5, Rank-20 scores to represent the CMC curve.

Initial data selection. To initialize the labeled and unlabeled sets, we follow the protocol outlined in [30]. For each identity, a tracklet is chosen randomly in camera 1. If camera 1 does not record an identity, a tracklet in the next available camera is chosen to ensure each identity has one tracklet for initialization.

Implementation details. Please see supplementary material for details on implementation, values of different hyper-parameters and datasets.

Comparison to the State-Of-The-Art methods. One-shot re-ID methods in the literature can be broadly divided into two classes: (1) DGM [32] and Stepwise Metric [18] use the entire pseudo-labeled data at each step of learning and in the process incorporate a lot of noisy labels, (2) EUG [30] and One-Example Progressive Learning [29] employ progressive sampling. TCPL outperforms all of these by learning an embedding which is temporally consistent. We also consider two baselines: Baseline (one-shot), which utilizes only the one-shot data for training, and Baseline(supervised), which assumes all the tracklets in the training set are labeled; these are trained in a supervised manner using only the cross-entropy loss. We also compare against state-of-the-art unsupervised methods which report results on video re-ID datasets: BUC [16], UTAL [15] and DAL [6].

We present the results for different instantiations of our framework in Table 1: one which uses both the losses (TCPL -full) and two others corresponding to usage of the losses individually (TCPL -$\mathcal{L}_{\text{intra}}$,TCPL -$\mathcal{L}_{\text{inter}}$). For TCPL, EUG

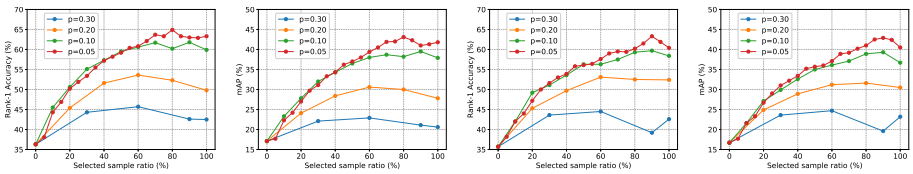
Table 1. Comparison of TCPL with state-of-the-art one-shot and unsupervised methods on the MARS and DukeMTMC-VideoReID datasets. (Sup./Unsup. refers to supervised and unsupervised methods respectively.)

Method	Setting	MARS			Duke		
		R-1	R-5	mAP	R-1	R-5	mAP
Baseline: upper bound	Sup.	80.8	92.1	67.4	83.6	94.6	78.3
TCPL -full (Ours)	1-shot	**65.2**	**77.5**	**43.6**	**76.8**	**87.8**	**67.9**
TCPL -$\mathcal{L}_{\text{intra}}$ (Ours)	1-shot	63.3	75.2	42.9	76.2	87.6	67.7
TCPL -$\mathcal{L}_{\text{inter}}$ (Ours)	1-shot	64.9	77.5	43.1	74.4	86.6	66.5
One-Shot Prog. [29]	1-shot	62.8	75.2	42.6	72.9	84.3	63.3
EUG [30]	1-shot	62.7	72.9	42.5	72.8	84.2	63.2
Stepwise Metric [18]	1-shot	41.2	55.6	19.7	56.3	70.4	46.8
DGM+IDE [32]	1-shot	36.8	54.0	16.9	42.4	57.9	33.6
Baseline: lower bound	1-shot	36.2	50.2	15.5	39.6	56.8	33.3
BUC [16]	Unsup.	61.1	75.1	38.0	69.2	81.1	61.9
UTAL [15]	Unsup.	49.9	66.4	35.2	-	-	-
DAL [6]	Unsup.	46.8	63.9	21.4	-	-	-

[30] and One-Shot Progressive [29], we set the enlarging parameter p to 0.05.The consistency losses lead to consistent gains of in both rank-1 accuracy and mAP over both EUG [30] and One-Shot Progressive Learning [29] in both the datasets.

Analysis over enlarging factor p. The selection of the enlarging factor p plays an important role in progressive sampling methods. Decreasing the value of p generally leads to less label estimation errors due to careful data selection, at the cost of a very slow learning process (See Fig. 3).

The performance of our method as p varies is shown in Table 2. Unlike baseline methods, which suffer drastic drops in performance as p is increased, our framework limits label estimation errors via the consitency losses. Notably, TCPL



(a) (b) (c) (d)

Fig. 3. Comparison with different values of enlarging factor on MARS. Figures (a) and (b) represent the Rank-1 accuracy and mAP using TCPL with $\mathcal{L}_{\text{inter}}$. Figures (c) and (d) represent the Rank-1 accuracy and mAP using TCPL with $\mathcal{L}_{\text{intra}}$.

Table 2. Variation in one-shot performance results for different scales of the enlarging parameter p. The best and second best results are in red/blue respectively.

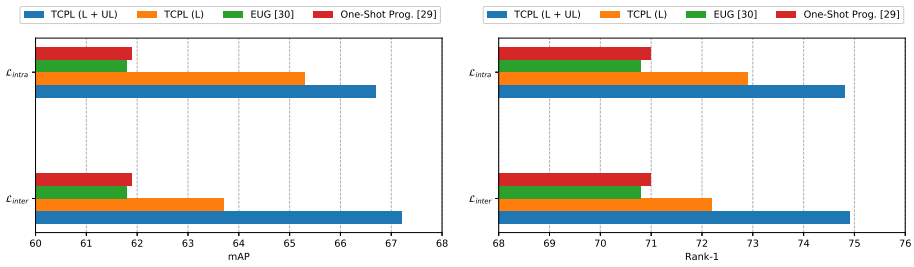
p	Method	Duke				MARS			
		R-1	R-5	R-20	mAP	R-1	R-5	R-20	mAP
0.20	EUG [30]	68.9	81.1	89.4	59.5	48.7	63.4	72.6	26.6
	One-Shot Prog. [29]	69.1	81.2	89.6	59.6	49.6	64.5	74.4	27.2
	TCPL -$\mathcal{L}_{\text{intra}}$	74.4	85.8	91.6	65.4	52.5	65.6	73.9	31.6
	TCPL -$\mathcal{L}_{\text{inter}}$	69.4	81.6	88.5	60.5	53.6	66.2	74.9	30.6
0.10	EUG [30]	70.8	83.6	89.6	61.8	57.6	69.6	78.1	34.7
	One-Shot Prog. [29]	71.0	83.8	90.3	61.9	57.9	70.3	79.3	34.9
	TCPL -$\mathcal{L}_{\text{intra}}$	74.8	87.3	92.0	66.7	59.7	72.0	79.3	39.3
	TCPL -$\mathcal{L}_{\text{inter}}$	74.9	86.5	92.0	67.2	61.8	74.7	81.5	39.5
0.05	EUG [30]	72.8	84.2	91.5	63.2	62.7	72.9	82.6	42.5
	One-Shot Prog. [29]	72.9	84.3	91.4	63.3	62.8	75.2	83.8	42.6
	TCPL -$\mathcal{L}_{\text{intra}}$	76.2	87.6	92.9	67.7	63.3	75.2	82.4	42.9
	TCPL -$\mathcal{L}_{\text{inter}}$	74.4	86.6	92.2	66.5	64.9	77.5	84.1	43.1

at $p = 0.20$ is able to outperform both EUG and One-Shot Progressive Learning at $p = 0.05$ on DukeMTMC-VideoReID. This translates to a **4**× speedup of learning without sacrificing performance. On MARS, at $p = 0.10$, TCPL is able to achieve a Rank-1 accuracy of 61.8%. This is only 1% behind One-Shot Progressive Learning with $p = 0.05$ and suggests a **2**× speedup with only a negligible drop in performance. All of these indicate that TCPL is robust to appending pseudo-labeled data more aggressively and thus, can save time.

Importance of maximally using the unlabeled data. The ability to extract maximal information from the unlabeled data is at the core of TCPL. We demonstrate this in Fig. 4 by evaluating the losses on DukeMTMC-VideoReID with and without access to entire unlabeled data at each step of learning.

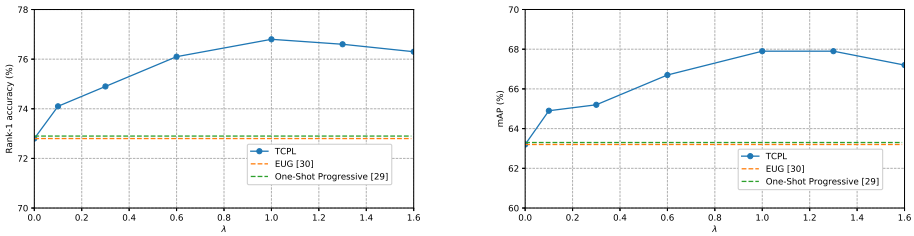
The results confirm the two aspects of our hypothesis. Firstly, utilizing the entire unlabeled set at every step of learning improves performance. Secondly, self-supervision - even without access to the entire unlabeled set - learns better features and improves re-ID performance. TCPL, with access to only the labeled data, outperforms [29] which accesses the entirety of the unlabeled set. This is a direct consequence of the ability of self-supervision to learn better features via consistency regularization, within and across camera views.

Weight on the loss function. In our framework, we jointly optimize two types of losses - the cross-entropy loss and the temporal coherence losses ($\mathcal{L}_{\text{intra}}, \mathcal{L}_{\text{inter}}$), as defined in Eqn. 7, to learn the weights θ of the feature embedding $f_\theta(\cdot)$. We investigate the contributions of the temporal losses to the re-identification



(a) (b)

Fig. 4. Performance of TCPL by varying access to the unlabeled set. (a) presents the Rank-1 acc. and (b) the mAP on DukeMTMC-VideoReID. Temporal consistency performs better than [30, 29] without using the entire unlabeled data, and improves even further when the unlabeled data is used. This demonstrates two things: (1) using the unlabeled data efficiently is important, (2) self-supervision can learn highly discriminative features. (L/UL denote the labeled/unlabeled set.)



(a) (b)

Fig. 5. Importance of temporal consistency. (a) presents variations in Rank-1 accuracy on DukeMTMC-VideoReID by changing weights on temporal losses. Higher λ represents more weight on the temporal losses. (b) presents the variations in mAP.

performance. In order to do that, we performed experiments with different values of λ (higher value indicates larger weight on the temporal losses) and present the results on the DukeMTMC-VideoReID dataset in Fig. 5. In general, increasing the weight improves performance, indicating the efficacy of self-supervision. As may be observed from the plot, the proposed method performs best with $\lambda = 1$.

Analysis over pseudo-label estimation. As a consequence of more discriminative feature learning using local consistency, TCPL is able to generate high quality labels for the unlabeled set. At $p = 0.20$ and $p = 0.10$, TCPL is able to achieve **8.2**% and **4.0**% improvement in label estimation respectively, on DukeMTMC-VideoReID, compared to EUG. On MARS, the improvement in estimation is **5.0**% and **3.8**% respectively. A visual representation of the improved pseudo-label estimation can be found in Fig. 6.

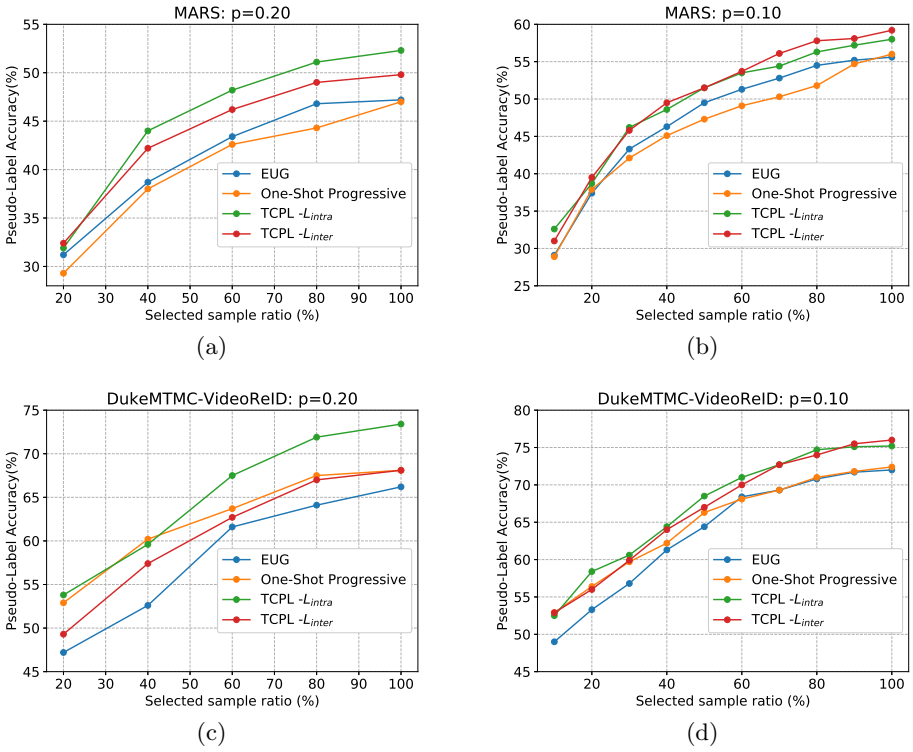


Fig. 6. Pseudo-label estimation. Accuracy of pseudo-labels as enlarging factor p is varied, on MARS [(a), (b)] and DukeMTMC-VideoReID [(c), (d)]

5 Conclusion

In this paper, we introduce a new framework, Temporally Consistent Progressive Learning, which uses self-supervision via temporal coherence, in conjunction with one-shot labels, to learn a person re-ID model. Two novel temporal consistency losses, intra-sequence temporal consistency and inter-sequence temporal consistency, are at the core of this framework. These losses enable learning of richer and more discriminative representations. Our approach demonstrates the importance of using the unlabeled data efficiently and intelligently, an aspect of one-shot re-ID ignored by most previous works. Experiments on two challenging datasets establish our method as the state-of-the-art in the one-shot video person re-ID task. Future work will concentrate on extending the idea of temporal coherence to unsupervised person re-identification.

Acknowledgments. We thank Sourya Roy, Sujoy Paul and Abhishek Aich for their assistance, advice and critique. The work was partially supported by NSF grant 1544969 and ONR grant N00014-19-1-2264.

References

1. Arazo, E., Ortego, D., Albert, P., O'Connor, N.E., McGuinness, K.: Pseudo-labeling and confirmation bias in deep semi-supervised learning. arXiv preprint arXiv:1908.02983 (2019)
2. Bak, S., Carr, P.: One-shot metric learning for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2990–2999 (2017)
3. Chapelle, O., Scholkopf, B., Zien, A.: Semi-supervised learning. IEEE Transactions on Neural Networks **20**(3), 542–542 (2009)
4. Chen, D., Li, H., Xiao, T., Yi, S., Wang, X.: Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1169–1178 (2018)
5. Chen, T., Ding, S., Xie, J., Yuan, Y., Chen, W., Yang, Y., Ren, Z., Wang, Z.: Abd-net: Attentive but diverse person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 8351–8361 (2019)
6. Chen, Y., Zhu, X., Gong, S.: Deep association learning for unsupervised video person re-identification. In: Proceedings of the British Machine Vision Conference (2018)
7. Ding, G., Zhang, S., Khan, S., Tang, Z., Zhang, J., Porikli, F.: Feature affinity based pseudo labeling for semi-supervised person re-identification. IEEE Transactions on Multimedia (2019)
8. Dosovitskiy, A., Springenberg, J.T., Riedmiller, M., Brox, T.: Discriminative unsupervised feature learning with convolutional neural networks. In: Advances in Neural Information Processing Systems. pp. 766–774 (2014)
9. Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. arXiv preprint arXiv:1803.07728 (2018)
10. Hamid Rezatofighi, S., Milan, A., Zhang, Z., Shi, Q., Dick, A., Reid, I.: Joint probabilistic matching using m-best solutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 136–145 (2016)
11. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 (2017)
12. Kolesnikov, A., Zhai, X., Beyer, L.: Revisiting self-supervised visual representation learning. arXiv preprint arXiv:1901.09005 (2019)
13. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems. pp. 1097–1105 (2012)
14. Lee, D.H.: Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: Workshop on Challenges in Representation Learning, ICML. vol. 3 (2013)
15. Li, M., Zhu, X., Gong, S.: Unsupervised tracklet person re-identification. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019)
16. Lin, Y., Dong, X., Zheng, L., Yan, Y., Yang, Y.: A bottom-up clustering approach to unsupervised person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 8738–8745 (2019)
17. Liu, X., Song, M., Tao, D., Zhou, X., Chen, C., Bu, J.: Semi-supervised coupled dictionary learning for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3550–3557 (2014)

18. Liu, Z., Wang, D., Lu, H.: Stepwise metric promotion for unsupervised video person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2429–2438 (2017)
19. Misra, I., Zitnick, C.L., Hebert, M.: Shuffle and learn: unsupervised learning using temporal order verification. In: Proceedings of the European Conference on Computer Vision. pp. 527–544. Springer (2016)
20. Miyato, T., Maeda, S.i., Koyama, M., Ishii, S.: Virtual adversarial training: a regularization method for supervised and semi-supervised learning. IEEE transactions on Pattern Analysis and Machine Intelligence **41**(8), 1979–1993 (2018)
21. Mobahi, H., Collobert, R., Weston, J.: Deep learning from temporal coherence in video. In: Proceedings of the 26th Annual International Conference on Machine Learning. pp. 737–744. ACM (2009)
22. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: Proceedings of the European Conference on Computer Vision. pp. 69–84. Springer (2016)
23. Oliver, A., Odena, A., Raffel, C.A., Cubuk, E.D., Goodfellow, I.: Realistic evaluation of deep semi-supervised learning algorithms. In: Advances in Neural Information Processing Systems. pp. 3235–3246 (2018)
24. Paul, S., Roy, S., Roy-Chowdhury, A.K.: Incorporating scalability in unsupervised spatio-temporal feature learning. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 1503–1507. IEEE (2018)
25. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems. pp. 91–99 (2015)
26. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: Proceedings of the European Conference on Computer Vision. pp. 17–35. Springer (2016)
27. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: Advances in Neural Information Processing Systems. pp. 1195–1204 (2017)
28. Wang, G., Lai, J., Huang, P., Xie, X.: Spatial-temporal person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 8933–8940 (2019)
29. Wu, Y., Lin, Y., Dong, X., Yan, Y., Bian, W., Yang, Y.: Progressive learning for person re-identification with one example. IEEE Transactions on Image Processing **28**(6), 2872–2881 (2019)
30. Wu, Y., Lin, Y., Dong, X., Yan, Y., Ouyang, W., Yang, Y.: Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5177–5186 (2018)
31. Xu, S., Cheng, Y., Gu, K., Yang, Y., Chang, S., Zhou, P.: Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4733–4742 (2017)
32. Ye, M., Ma, A.J., Zheng, L., Li, J., Yuen, P.C.: Dynamic label graph matching for unsupervised video re-identification. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5142–5150 (2017)
33. Ye, M., Zhang, X., Yuen, P.C., Chang, S.F.: Unsupervised embedding learning via invariant and spreading instance feature. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6210–6219 (2019)

34. Yu, H.X., Wu, A., Zheng, W.S.: Cross-view asymmetric metric learning for unsupervised person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 994–1002 (2017)
35. Yu, H.X., Zheng, W.S., Wu, A., Guo, X., Gong, S., Lai, J.H.: Unsupervised person re-identification by soft multilabel learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2148–2157 (2019)
36. Zheng, L., Bie, Z., Sun, Y., Wang, J., Su, C., Wang, S., Tian, Q.: Mars: A video benchmark for large-scale person re-identification. In: Proceedings of the European Conference on Computer Vision. pp. 868–884. Springer (2016)
37. Zheng, L., Yang, Y., Hauptmann, A.G.: Person re-identification: Past, present and future. arXiv preprint arXiv:1610.02984 (2016)
38. Zhou, K., Yang, Y., Cavallaro, A., Xiang, T.: Omni-scale feature learning for person re-identification. arXiv preprint arXiv:1905.00953 (2019)