Neural Networks 129 (2020) 43–54

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

An end-to-end exemplar association for unsupervised person Re-identification

Jinlin Wu [a,b,1], Yang Yang [a,b,1], Zhen Lei [a,b,*], Jinqiao Wang [a,b], Stan Z. Li [c], Prayag Tiwari [d], Hari Mohan Pandey [e]

[a] CBSR & NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[b] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[c] School of Engineering, Westlake University, Hangzhou, China
[d] Department of Information Engineering, University of Padova, Italy
[e] Department of Computer Science, Edge Hill University, Ormskirk, United Kingdom

ARTICLE INFO

Article history:
Received 3 January 2020
Received in revised form 19 April 2020
Accepted 11 May 2020
Available online 23 May 2020

Keywords:
End-to-end exemplar-based training
Exemplar association
Dynamic selection threshold

ABSTRACT

Tracklet association methods learn the cross camera retrieval ability though associating underlying cross camera positive samples, which have proven to be successful in unsupervised person re-identification task. However, most of them use poor-efficiency association strategies which costs long training hours but gains the low performance. To solve this, we propose an effective end-to-end exemplar associations (EEA) framework in this work. EEA mainly adapts three strategies to improve efficiency: (1) **end-to-end exemplar-based training**, (2) **exemplar association** and (3) **dynamic selection threshold**. The first one is to accelerate the training process, while the others aim to improve the tracklet association precision. Compared with existing tracklet associating methods, EEA obviously reduces the training cost and achieves the higher performance. Extensive experiments and ablation studies on seven RE-ID datasets demonstrate the superiority of the proposed EEA over most state-of-the-art unsupervised and domain adaptation RE-ID methods.

1. Introduction

Person Re-identification (RE-ID) is an open-set retrieval task and has potential applications like longterm multi-camera tracking and forensic search. Given a pedestrian images, the person Re-ID model encodes it to a representation vector and uses the representation vector retrieving similar pedestrian images across different cameras. Essentially, person RE-ID is a representation learning task which learns the view-invariant visual representation. Benefiting from the advance of deep representation learning (He, Zhang, Ren, & Sun, 2016; Krizhevsky, Sutskever, & Hinton, 2012) and deep metric learning (Hadsell, Chopra, & LeCun, 2006; Liu, Zhu, Lei, & Li, 2019; Schroff, Kalenichenko, & Philbin, 2015), the performance of RE-ID has obtained significant improvements (Bai, Bai, & Tian, 2017; Hou et al., 2019b; Li, Zhao, Xiao, & Wang, 2014a; Shen, Li, Yi, Chen, & Wang, 2018; Shi et al., 2016; Sun, Zheng, Deng, & Wang, 2017; Sun, Zheng, Yang, Tian, & Wang, 2018; Yi, Lei, Liao, & Li, 2014; Zheng, Zheng, & Yang, 2018). These deep person RE-ID methods are data-driven supervised algorithms, which need a large number of pair-wise labeled data to learn the view-invariant representations. Fig. 1 shows examples of pair-wise labeled tracklets. Pair-wise labeling denotes that annotating the some pedestrian from different cameras, which is expensive and time-consuming. Hence, unsupervised training and improving the scalability of deep RE-ID algorithm become the great challenges in recent person RE-ID research.

There have been a series of unsupervised image-level methods to address this problem, which can be roughly divided into four categories: (1) image style transformation, (2) model domain adaptation, (3) unsupervised clustering and (4) memory association. **Image style transformation** methods (Bak, Carr, & Lalonde, 2018; Deng, Zheng, Ye, Kang, Yang, & Jiao, 2018; Wei, Zhang, Gao, & Tian, 2018; Zhong, Zheng, Li and Yang, 2018; Zhong, Zheng, Zheng, Li and Yang, 2018) transfer the source domain images to the target domain by GAN (Goodfellow et al., 2014; Zhu, Park, Isola, & Efros, 2017) network and train the target model with transferred images. **Domain adaptation** methods (Li et al., 2018; Wang, Zhu, Gong, & Li, 2018) aim to transfer the knowledge of the source domain trained model to the target domain in an unsupervised manner. **Clustering** methods (Fan, Zheng, Yan, & Yang, 2018; Jinlin, Shengcai, Zhen, Xiaobo, Yang, & Li, 2018)

[*] Corresponding author at: CBSR & NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing, China.

E-mail addresses: jinlin.wu@nlpr.ia.ac.cn (J. Wu), yang.yang@nlpr.ia.ac.cn (Y. Yang), zlei@nlpr.ia.ac.cn (Z. Lei), jqwang@nlpr.ia.ac.cn (J. Wang), szli@nlpr.ia.ac.cn (S.Z. Li), prayag.tiwari@dei.unipd.it (P. Tiwari), pandeyh@edgehill.ac.uk (H.M. Pandey).

[1] Equal contribution.

https://doi.org/10.1016/j.neunet.2020.05.015
0893-6080/© 2020 Elsevier Ltd. All rights reserved.

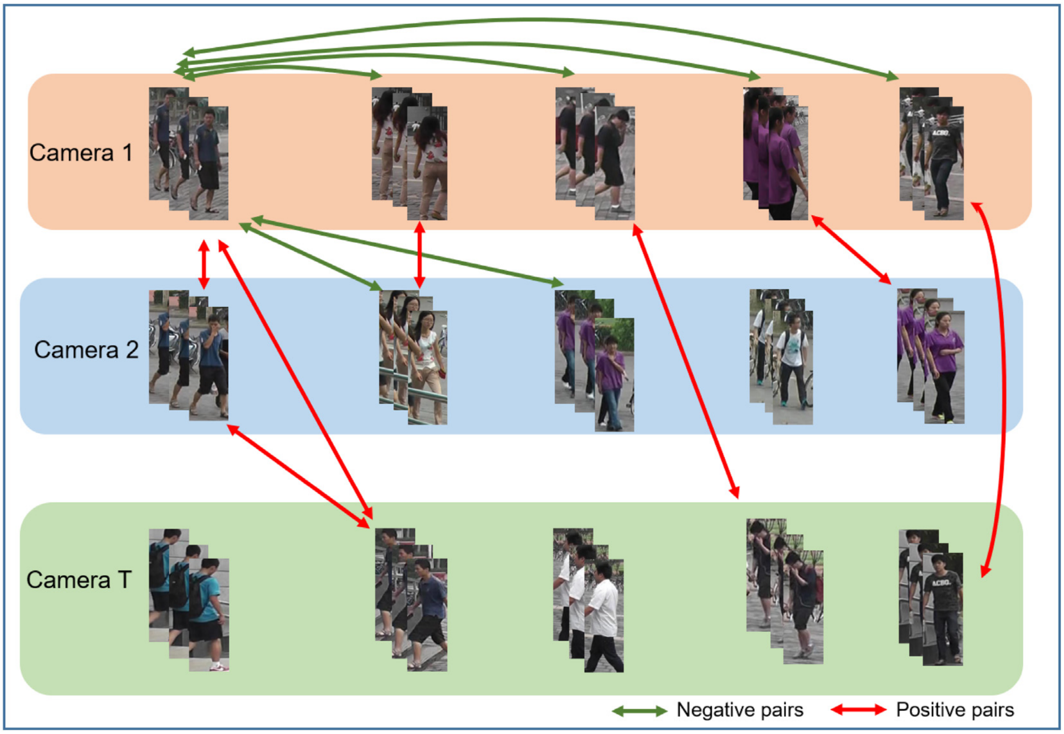


Fig. 1. Examples of the pair-wise labeled tracklets. Pair-wise labeled tracklets refer the images belonging to the same person under different cameras. RE-ID model learns the view-invariance representation by pull the positive pairs close and push the negative pairs away.

obtain pseudo labels of the target domain data through clustering algorithms and fine tune the source domain model with these pseudo labels. **Memory association methods** (Yu et al., 2019; Zhong, Zheng, Luo, Li and Yang, 2019) store the image feature in the memory and learn, building relations among image memories. These relations provide rich training signals in the unsupervised training process.

However, the precondition of above mentioned methods is domain similarity between the source domain and the target domain, since all of them need source domain pre-training. For example, as shown in Deng et al. (2018), Fan et al. (2018) and Liang, Wang, Lai, and Zhu (2018), the above mentioned methods are easily achieving high improvements on Market1501 (Zheng et al., 2015) and DukeMTMC-ReID (Ristani, Solera, Zou, Cucchiara, & Tomasi, 2016a). This is because Market1501 and DukeMTMC-ReID are similar to each other and the source domain pre-trained model can provide a high start point for unsupervised algorithms. While the pre-trained model hardly works on target datasets (*i.e.*, MSMT17, CUHK03), these unsupervised methods are difficult to play a useful role. This weakness limits the scalability of the above unsupervised algorithms for the real-world unknown scene applications.

Recent advanced methods overcome this weakness by using tracklet association methods (*i.e.*, TAUDL (Li, Zhu and Gong, 2018), UTAL (Li, Zhu, & Gong, 2019), RACE (Ye, Lan, & Yuen, 2018), BUC (Lin, Dong, Zheng, Yan, & Yang, 2019), UGA (Wu et al., 2019)). They assume that pedestrian tracklets are automatically obtained by existing detection (Dollar, Wojek, Schiele, & Perona, 2012; Zhang, Benenson, & Schiele, 2017; Zhang, Wen, Bian, Lei, & Li, 2018) and tracking algorithms (Lealtaixe, Milan, Reid, Roth, & Schindler, 2015; Ristani, Solera, Zou, Cucchiara, & Tomasi, 2016b). Though Sparse Space–Time Tracklet (SSTT) sampling (more details are available in), duplicate tracklets can be removed which means that each person has at most one tracklet in each camera. There are no positive pairs in the same camera. Based on this assumption, these methods focus on mining cross camera positive tracklet pairs from cross camera retrieval ability learning. More details about SSTT are available in TAUDL (Li, Zhu et al., 2018) and UTAL (Li et al., 2019). However, the proposed association strategies of the above methods are not efficient enough which cost long training hours and large GPU memories but gain the low performance. The inefficiency reasons of RACE and BUC are they adapt the progressively tracklets merging strategy in training which is easily misled by merging noisy pairs and BUC has to take a very long time on tracklets clustering and merging. UTAL and TAUDL propose a multi-camera-branch to learn intra-camera representations and an on-line association strategy to mining the underlying positive pairs. But the on-line association is poorly efficient because it needs a large batch size (384) to sample underlying positive pairs and long training hours for algorithm converging which means it may occupy at least five 1080-Ti GPUs and cost over 200 epochs for training. UGA adapts a graph association strategy to alleviate this weakness, but its two-stage training still needs long hours (about 160 epochs) for training. To improve the effective, we propose an end-to-end framework EEA. It can be simply implemented within 80 epochs with one 1080-Ti GPU. The pip-line of EEA is shown in Fig. 2, which mainly contains three strategies **end-to-end exemplar-based training**, **exemplar graph association** and (3) **dynamic selection threshold**, as described below.

End-to-end exemplar-based training framework. The deep person RE-ID can be regarded as a deep visual representation learning task. Similarly, the unsupervised person RE-ID task can be regarded as an unsupervised visual representation learning task. Inspired by existing unsupervised discriminative representation learning methods Exemplar-CNN (Dosovitskiy, Fischer, Springenberg, Riedmiller, & Brox, 2015) and Momentum (He, Fan, Wu, Xie, & Girshick, 2019), we adapt an exemplar-based method in this study. We regard each tracklet as a single exemplar and design an exemplar memory module to store it. However, different with Exemplar-CNN and Momentum, RE-ID is an across camera retrieval task instead of image recognition. Taking this

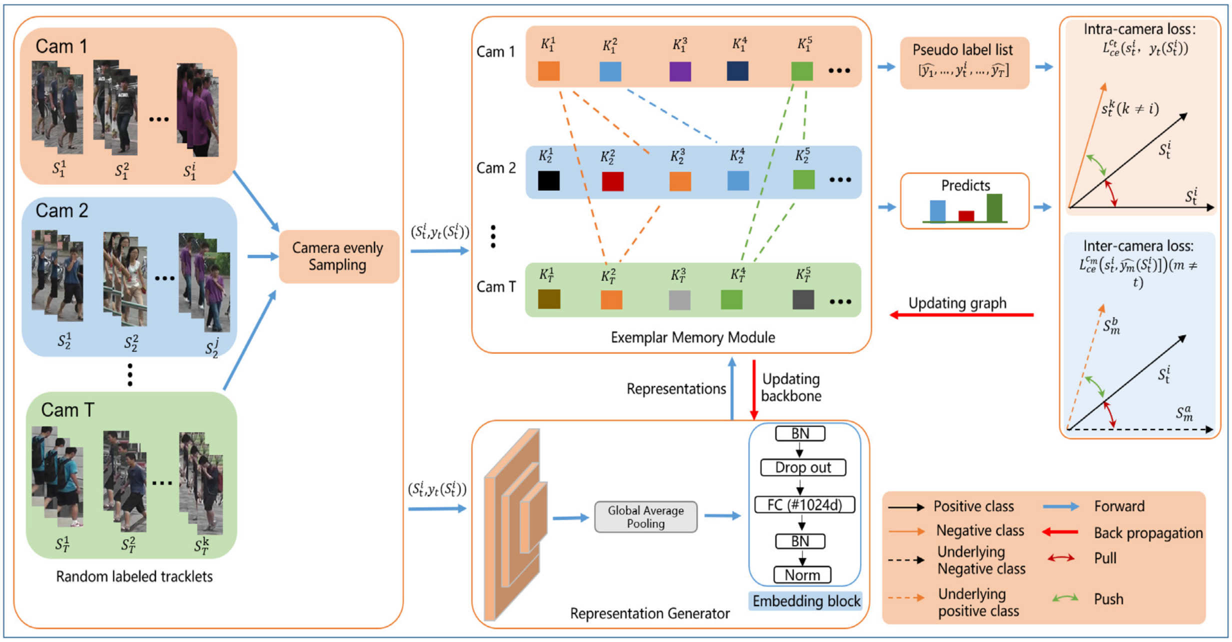


Fig. 2. The pip-line of the end-to-end exemplar association framework (EEA), including (1) a representation generator and (2) an exemplar memory module. The representation generator consists of a Resnet-50 backbone, a global average pooling layer (GAP), an embedding block. This embedding block contains a batch normalization layer, a drop-out layer, a FC layer reducing the 2048-dim feature to 1024-dim, a batch normalization layer and a L_2 normalization layer. The exemplar module is used as surrogate classifier storing the representation of each tracklet exemplar.

character into account, we investigate unsupervised RE-ID representation learning from intra-camera representations learning and inter-camera representations learning. For the former, we push different exemplar away to learn distinguishing different people. While for the latter, the underlying positive exemplar pair should be pulled close to learn retrieving the same pedestrian across different cameras. An example of these two representation learning is shown in Fig. 1 and the framework is shown in Fig. 2. In this framework, the backbone is used as a representation generator. The exemplar memory module is used as the surrogate classifier for intra-camera loss and inter-camera loss. It allows EEA simultaneously learning the intra-camera representation and the inter-camera representation. Furthermore, the exemplar memory module be directly updated by inter-camera loss and intra-camera loss in the back propagation. Comparing with off-line updating (*i.e.,* ECN Zhong, Zheng, Luo et al. (2019) and BUC Lin et al. (2019)), it improves the training speed.

Exemplar association. Most existing works simply apply KNN graph for tracklet associations. However KNN graph is highly computationally complex and includes many noisy associations. In order to alleviate these weakness, we build a more accurate graph cross-camera graph on the exemplar memory module. Three constraints (cross-camera, threshold, symmetry) are used to reduce the noisy associations in the graph building process. The cross-camera constrain reduces the computational complexity, while the others improve the association correct.

Dynamic selection threshold The threshold used in cross-camera graph is very cumbersome to set manually. Especially different datasets almost have different suitable thresholds. If the threshold is setted too large, many truth positive pairs are missed. On the contrary, many noisy pairs are introduced in the graph. In this study, we adapt a dynamic selection threshold, which can be changed in the training process and has a wider range for selecting training samples.

To sum up, the contributions of this paper can be summarized as follows:

- We propose an efficient end-to-end unsupervised person RE-ID framework, named end-to-end exemplar association (EEA). Without any source domain pre-training, EEA achieves high performance, with low training costs (lower GPU occupation and faster training speed).
- In EEA, an exemplar memory module and an exemplar association strategy are proposed for view-invariant representation learning. The former can be fast updated in the back propagation and the latter is more efficient than KNN graph.
- We conduct extensive experiments and ablation studies on seven RE-ID datasets to demonstrate the effectiveness of the proposed EEA.

2. Related work

Deep supervised person RE-ID. The aim of person re-identification (RE-ID) is retrieving the same person under multiple views. Benefitting from the advance of the deep learning (Schmidhuber, 2015; Tavanaei, Ghodrati, Kheradpisheh, Masquelier, & Maida, 2018) algorithm, person RE-ID has achieved a remarkable progress (Chang, Hospedales, & Xiang, 2018; Shen et al., 2018; Sun et al., 2018; Tan et al., 2019; Wang, Lai, Huang, & Xie, 2019; Wang, Yang, Cheng, Wang, & Hou, 2019; Yi et al., 2014; Zheng et al., 2018). Yi et al. (2014) adopt image pairs and introduce part priors into a siamese network for learning the view-invariant representations. Chang et al. (2018) and Sun et al. (2018) develop the part feature based methods to enhance the discriminative of Re-ID features. Wang et al. (2019) fuse the temporal–spatial information with appearance information to improve the retrieval accuracy.

Unsupervised person RE-ID. Deep person RE-ID algorithm has poor scalability in real-word applications, due to the lack of sufficient pair-wise labeled data for training. To solve this problem, lots of unsupervised person RE-ID methods are proposed (Fan et al., 2018; Li, Yang et al., 2018; Wang et al., 2018; Zheng, Zheng, & Yang, 2017; Zhong, Zheng, Li et al., 2018; Zhong, Zheng, Zheng et al., 2018; Zhong, Zheng, Zheng, Li and Yang, 2019). Bak et al. (2018), Deng et al. (2018), Zhong, Zheng, Li et al. (2018) and Zhong, Zheng, Zheng et al. (2018) adopt the GAN network to transfer the source domain training images to target domains, or transfer the target domain testing images to the source domain for improving the testing accuracy. Li, Yang et al. (2018) and Wang, Gong, Zhu, and Wang (2014) apply the domain adaptation methods transferring source domain knowledge to target domain. Fan et al. (2018) and Wu et al. (Jinlin et al., 2018) fine tune the source model in target domain with target data pseudo labels, which are obtained by the unsupervised clustering algorithm. However, these methods rely on the similarity between the source domain and the target domain. In order to reduce the dependence on the source domain, the tracklet-based methods are proposed. Li, Zhu et al. (2018) and Li et al. (2019) match the underlying positive pairs in the mini batch, using a cross camera histogram loss to learn the view-invariant features. Ye et al. (2018) propose a robust embedding to reduce the damage of the noisy frames for estimator pseudo labels more accuracy. It largely closes the gap between unsupervised and supervised representation learning in many computer vision tasks.

Unsupervised representation learning. Unsupervised representation learning from visual data is long research hotspot in computer vision. Many famous models are proposed to solve this task, *i.e.*, auto-encoder (Bourlard & Kamp, 1988; Hinton & Salakhutdinov, 2006; Makhzani, Shlens, Jaitly, Goodfellow, & Frey, 2015), GAN (Goodfellow et al., 2014). Auto-encoder learns the mid-level representation though reconstructing images. GAN proposes the adversarial training to learn the visual data distribution. Recent years, Exemplar-CNN (Dosovitskiy et al., 2015) and Momentum are proposed for unsupervised representation learning. Exemplar-CNN randomly uses image patch as surrogate class and uses it to train a discriminative CNN model. Momentum (He et al., 2019) regards each image as an exemplar, proposing a contrastive loss and shuffling batch normalization trick to train a powerful pre-trained model for downstream computer vision task *i.e.*, detection, segmentation.

Graph association methods. Considering the relationships between the training samples, graph association methods (Chen, Xu, Li, Sebe and Wang, 2018; Gong et al., 2015; Luo, Zhu, Li, Ren, & Zhang, 2018; Shen et al., 2018) are used to provide more supervision signals for both of semi-supervised learning and supervised training. Luo et al. (2018) propose a smoothing neighbors on teacher loss (SNTG) for semi-supervised learning. SNTG builds the relation graph of training samples and learns more smoothing representations from the relation graph. SNTG is a semi-supervised method, which deals the closed set classification and needs a few of labeled samples for training. However, it is not suitable for the unsupervised person RE-ID task, since unsupervised person RE-ID is an open-set retrieval problem. Shen et al. (2018) propose a similarity-guided graph neural network (SGGNN) to enhance the relations between the probe images and the gallery pedestrian images. But SGGNN is a supervised training approach which needs lots of labeled samples to build the graph for training.

3. Method

3.1. Overview of EEA

Definition. Suppose we have a unlabeled dataset captured from T cameras and apply a Sparse Space–time Tracklets sampling (SSTT) algorithm to sample M tracklets for training. Denoting $s_t^i = \{I_1^{s_t^i}, I_2^{s_t^i}, \dots, I_n^{s_t^i}\}$, where $I_n^{s_t^i}$ is the nth image of the ith tracklet ($i \in [1, \dots, M_t]$) in tth camera ($t \in [1, \dots, T]$). A pseudo label list $[\hat{y_1}, \dots, y_t^i, \dots, \hat{y_T}]$ is given to the tracklet s_t^i for computing loss function. In the pseudo label list, y_t^i is used for intra-camera loss, which is randomly assigned under camera t before training and cannot be changed in training. While $\hat{y_m}(m \neq t)$ is applied for inter-camera loss, which is assigned by the cross-camera graph at the beginning of each training epoch. $\phi(\cdot)$ denotes the representation generator. $E_t^i (E_t^i \in E_t)$ is the exemplar stored in the exemplar memory module regarding as the proxy of s_t^i. E_t is a set of E_t^i used as a surrogate classifier for camera t.

Framework. EEA pip-line consists of three modules: a representation generator, an exemplar memory module and an camera evenly sampling module. The representation generator includes a Resnet-50 backbone, a global average pooling layer (GAP) and an embedding block. In which, the embedding block contains a batch normalization layer, a drop-out layer, a FC layer reducing the 2048-dim feature to 1024-dim, a batch normalization layer and a L_2 normalization layer. The exemplar module is used as surrogate classifier storing the representation of each tracklet exemplar.

3.2. Intra-camera representation learning

The intra-camera representation is learning from s_t^i and the corresponding pseudo label y_t^i. However, tracklets in different cameras belonging to the same person almost have different pseudo labels since y_t^i is random assigned in each camera. To avoid the conflict, we adopt the multi-task training to learn the intra-camera representation dependently, where each classifying task corresponds to a single camera training. All of the classifying tasks share the some representation generator. For the tth camera, the softmax cross-entropy loss function is formulated as follows:

$$l_{ce}^t(I_n^{s_t^i}, y_t^i, E_t) = -\sum_{j=1}^{M_t} \mathbf{1}(y_t^i == j) log(\frac{e^{(E_t^j)^T \phi(I_n^{s_t^i})}}{\sum_{k=1}^{M_t} e^{(E_t^k)^T \phi(I_n^{s_t^i})}}),$$
$$where\, E_t^j \in E_t \tag{1}$$

In Eq. (1), applying L2 normalization on E_t^j and $\phi(I_n^{s_t^i})$, the product of these two representations can be formulated as follows:

$$(E_t^j)^T \phi(I_n^{s_t^i}) = \|E_t^j\| \|\phi(I_n^{s_t^i})\| cos(\theta) = cos(\theta)$$
$$\times(\|E_t^j\| = 1, \|\phi(I_n^{s_t^i})\| = 1) \tag{2}$$

where θ is the angle between E_t^j and $\phi(I_n^{s_t^i})$. Suppose that there are two images, $I_1 (I_1 \in s_t^j)$ and $I_2 (I_2 \notin s_t^j)$, we can infer a result as follows:

$$cos(I_1, E_t^j) > cos(I_2, E_t^j) \tag{3}$$

The above equation is equal to that:

$$\|I_1 - E_t^j\| < \|I_2 - E_t^j\| \tag{4}$$

Eq. (4) indicates that the exemplars are centers of corresponding tracklets and can be learned by intra-camera loss. Due to this, we can directly learn the tracklet center from intra-camera representation learning and directly update it through back propagation. It is much faster than off-line updating which needs much time to re-generating representation and computing centers for the whole training set. The intra-camera loss l_{intra} of a mini batch can be defined as Eq. (5), where N_{bs} denotes the batch size.

$$l_{intra} = \frac{1}{N_{bs}} \sum_{N_{bs}} l_{ce}^t(I_n^{s_t^i}, y_t^i) \tag{5}$$

3.3. Inter-camera representation learning

Cross-camera graph building We define a local KNN set $\{E_t^i\}_K^m (m \neq t)$ of E_t^i on the exemplar memory module, which finds of the nearest K tracklets of E_t^i in camera m. There are no positive pairs in the same camera since SSTT sampling. In order to reduce

```
Algorithm 1: End-to-End exemplar association (EEA)
Input: Unlabeled tracklets $s_t^i$ of $T$ cameras.
       The representation generator $\phi(\cdot)$.
       An randomly initialized exemplar memory module $E$.
       $E_t (E_t \in E)$ is the exemplar set of camera $t$.
       Threshold $\lambda$ and max iteration $ep_{max}$.
       The lower bound $L(\lambda)$ and changing rate $\eta$.
       Exemplar initializing epoch $ep_{warm}$ for exemplars initiating.
       Randomly assigning pseudo labels $y_t^i$ for each tracklet.
       Initializing $ep \leftarrow 0$, $\lambda \leftarrow L(\lambda)$.
while $ep < ep_{max}$ do
    1: Evenly sampling tracklets from $T$ cameras;
    2: Assigning cross-camera pseudo list $[\hat{y_1}, \cdots, y_m^i, \cdots, \hat{y_T}]$ with
    exemplar memory module;
    3: if $ep > ep_{warm}$ then
        $\lambda \leftarrow \lambda + \eta$ ;
    end
    4: $t \leftarrow 0$ ;
    5: for  $t$ in $[0, \cdots, T]$ do
        Computing intra-camera loss $l_{intra}$ for camera $t$ by Eq. (5);
        Computing inter-camera loss $l_{inter}$ for other $T-1$ cameras by
        Eq. (9);
        Computing total loss $L_{total}$ according Eq. (10);
    end
    6: Updating $\phi$ and $E$;
    7: $ep \leftarrow ep + 1$;
end
Output: Representation generator $\phi$
.
```

computation complexity, the local KNN set only finds the nearest K tracklets between different cameras. Through merging these local KNN sets, we can build a cross-camera graph with M nodes. An adjacency matrix $A \in R^{M \times M}$ is used to denote the graph. The edge in the graph can be defined as:

$$A(E_t^i, E_m^j) = \begin{cases} cos(E_t^i, E_m^j), & if\ cos(E_t^i, E_m^j) > \lambda\ \&\ Sym(E_t^i, E_m^j) = True\ \&\ t \neq m \\ 1, & m = t\ \&\ i = j \\ 0, & else \end{cases} \tag{6}$$

In above equation, $cos(E_t^i, E_m^j) > \lambda$ is the threshold constrain, which requires the cosine similarity between the nodes E_t^i and E_m^j is larger than the threshold λ. $(E_t^i, E_m^j)_K$ is the symmetric constraint, which requires E_t^i and E_m^j must exist in each other's local KNN set. It can be defined as follows:

$$Sym(E_t^i, E_m^j) = \begin{cases} True, & E_m^j \in \{E_t^i\}_K^m\ \&\ E_t^i \in \{E_m^j\}_K^t\ \&\ t \neq m \\ False, & else \end{cases} \tag{7}$$

Considering through the SSTT sampling, each person has at most one tracklet in each camera. K is set to 1 in Eq. (6) and Eq. (7).

Comparing the conventional KNN graph, suppose that each camera has N tracklets and T cameras in total. KNN graph needs to compute a $TN \times TN$ matrix and sort $TN \times TN$ matrix to find top-K nearest neighbors. While our cross-camera graph computes and sorts $N \times N$ matrices for $\binom{2}{T}$ times. The $N \times N$ matrix is easier and faster to compute and sort than the $TN \times TN$ matrix. An intuitive explanation is that the cross-camera graph does not count tracklets pairs belonging to the same camera.

Inter-camera loss. Firstly, we define a graph neighbor set $N(s_t^i)$ of the tracklet s_t^i:

$$N(s_t^i) = \{(s_m^a, y_m^a) | A(E_t^i, E_m^a) \neq 0, m \in [1, \dots, T], m \neq t\} \tag{8}$$

We think these tracklets $s_m^a (s_m^a \in N(s_t^i))$ belonging to the same graph neighbor set are the underlying positive pairs. In fact, $N(s_t^i)$ is a set which contains s_t^i's all local nearest neighbors of other $T-1$ cameras. In other $T-1$ cameras, s_t^i may have same pseudo labels with its corresponding graph neighbor. Therefore, for the graph neighbors $[s_1^a, \dots, s_m^b, \dots, s_T^c] (s_m^b \in N(s_t^i))$ of s_t^i, we given the corresponding pseudo labels $[y_1^a, \dots, y_m^b, \dots, y_T^c]$ for training s_t^i in other $T-1$ cameras to learn inter-camera representations. In order to distinguish with the pseudo y_t^i, we simply denote $[y_1^a, \dots, y_m^b, \dots, y_T^c]$ as $[\hat{y_1}, \dots, \hat{y_m}, \dots, \hat{y_T}]$. To this end, we propose the following inter-camera loss to pulling these underlying positive pairs close:

$$\begin{aligned} l_{ce}^m(I_n^{s_t^i}, \hat{y_m}, E_m) &= -\sum_{j=1}^{M_m} \mathbf{1}(\hat{y_m} == j) log(\frac{e^{(c_m^j)^T \phi(I_n^{s_t^i})}}{\sum_{k=1}^{M_m} e^{(c_m^k)^T \phi(I_n^{s_t^i})}}) \\ l_{inter}(I_n^{s_t^i}) &= \sum_{N(s_t^i)} A(E_t^i, E_m^a) l_{ce}^m(I_n^{s_t^i}, \hat{y_m}, E_m) \\ &= \sum_{m=1, m\neq t}^{T} A(E_t^i, E_m^a) l_{ce}^m(I_n^{s_t^i}, \hat{y_m}, E_m), \\ & where\ \hat{y_m} = y_m^a \end{aligned} \tag{9}$$

In Eq. (9), edge weight $A(E_t^i, E_m^a)$ is used as a confidence coefficient to avoid misled by noisy associations. The total loss can be summarized as follows:

$$\begin{aligned} L_{total} &= l_{ce}^t(I_n^{s_t^i}, y_t^i, E_t) + \sum_{m=1, m\neq t}^{T} A(E_t^i, E_m^a) l_{ce}^m(I_n^{s_t^i}, \hat{y_m}, E_m) \\ L_{total} &= \sum_{m=1}^{T} A(E_t^i, E_m^a) l_{ce}^m(I_n^{s_t^i}, \hat{y_m}, E_m) \quad where\ A(E_t^i, E_m^a) = 1 \end{aligned} \tag{10}$$

In above equation, both of the intra-camera loss and the inter-camera loss can be computed on the exemplar memory. It allows us to adapt an end-to-end frame work to simultaneously learn intra-camera representation and inter-camera representation rather than adapt a multi-stage strategy to learn them.

3.4. Dynamic selection threshold

In last section, we introduce a cross-camera graph for inter-camera representation learning. It needs a threshold λ for training pairs selecting. We initiate the exemplar memory module with several epochs intra-camera training for investigating the influence of λ. As shown in Fig. 3, the small threshold has large recall and low precision which means it finds more correct pairs but introduces more noisy associations. The large threshold is the opposite. Due to this, we propose a dynamic changing strategy increasing λ from a lower bound $L(\lambda)$ to a upper bound $U(\lambda)$. In this study, we adapt a simple yet effective method, linearly increasing, which can be formulated as:

$$\begin{aligned} \lambda &=: \lambda + \eta \\ \eta &= \frac{U(\lambda) - L(\lambda)}{ep_{max} - ep_{warm}} \end{aligned} \tag{11}$$

In Eq. (11), ep_{max} denotes the total training epoch and ep_{warm} denotes the first several epochs for exemplar initializing.

4. Experiment

4.1. Experimental setup

Datasets and evaluation protocol. All experiments are evaluated on four image RE-ID datasets (Market-1501 (Zheng et al., 2015), DukeMTMC-ReID (Ristani et al., 2016a; Zheng et al., 2017), CUHK03-detected (Li et al., 2014b), MTMS17 (Wei et al., 2018)) and three video RE-ID datasets (Mars (Zheng et al.,

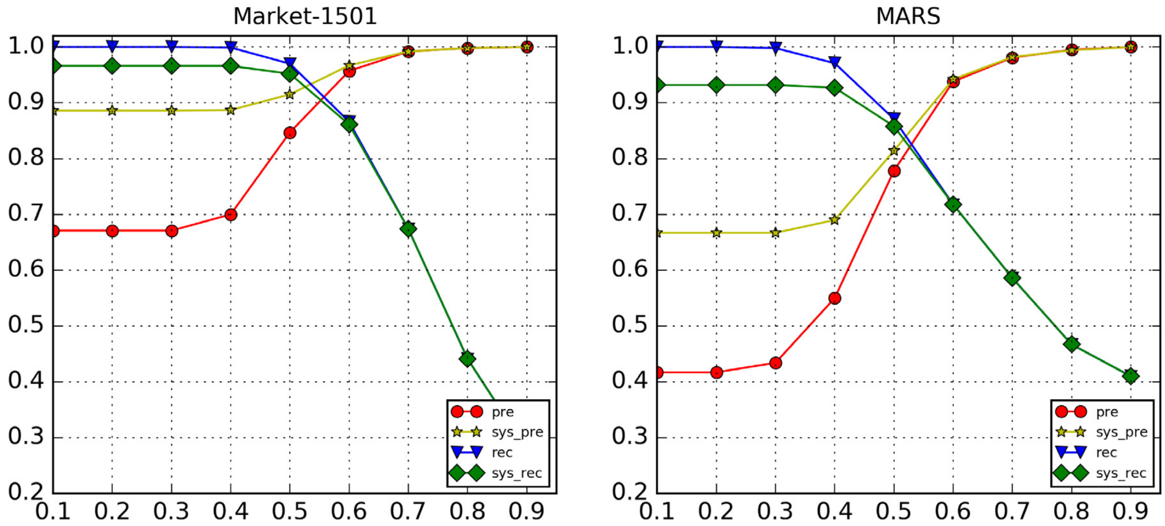


Fig. 3. Threshold analysis on Market-1501 and MARS. In this figure, "pre" denotes the precision. "rec" denotes the recall. "sys_pre" and "sys_rec" respectively denote the precision and recall of using the symmetry condition. The horizontal axis is the value of the λ. The vertical axis denotes the value of the precision score and recall score.

Table 1
Dataset statistics and training/testing splitting.

Dataset	ID	Cameras	Tracklets	Train	Test	Images
iLIDS-VID (Wang et al., 2014)	300	2	600	150	150	43,800
PRID2011 (Hirzer, Beleznai, Roth, & Bischof, 2011)	178	2	354	89	89	38,466
MARS (Zheng et al., 2016)	1261	6	20,478	625	636	1,191,003
Market-1501 (Zheng et al., 2015)	1501	6	–	751	750	32,668
DukeMTMC-ReID (Ristani et al., 2016a; Zheng et al., 2017)	1812	8	–	702	1110	36,411
MSMT17 (Wei et al., 2018)	4101	15	–	1041	3060	126,441
CUHK03-detected (Li, Zhao, Xiao, & Wang, 2014b)	1467	2	–	1367	100	14,096

Table 2
The ablation studies of the structure & sampling.

Strategies	Market-1501		Mars	
Metric (%)	mAP	Rank-1	mAP	Rank-1
*Res*50	12.0	28.9	20.9	34.9
*Res*50 + *KT*	27.9	47.3	26.0	41.1
*Res*50 + *KT* + *emb*	54.8	77.5	46.3	55.7

*Res*50 denotes only use the Resnet-50 backbone;
KT denotes camera evenly sampling;
emb denotes adding an embedding block at the top of the backbone.

2016), Prid2011 (Hirzer et al., 2011), iLIDS-Video (Wang et al., 2014)). The ablation studies are mainly conducted on Market-1501 (Zheng et al., 2015) and Mars (Zheng et al., 2016) which are most the widely used image and video person RE-ID datasets. The training/testing ID splits are shown in Table 1. Common cumulative matching characteristic (CMC) and mean average precision (mAP) are used as the performance evaluation metric. Particularly, on Market-1501, we follow the single-query evaluation protocol. On the CUHK03-detected, we follow the standard single-shot protocol for the labeled images and detected images separately, which needs to repeat 20 times of random 1367/100 training/testing identity splitting and report the averaged results.

Pseudo label assignment. We follow the experiments settings and tracklet sampling methods of TAUDL (Li, Zhu et al., 2018) and UTAL (Li et al., 2019). For video datasets, iLIDS-VID and PRID2011 provide only one tracklet of a person in one camera. But MARS has multiple tracklets per ID per camera. We randomly sample one tracklet for a person in one camera on MARS. For the image RE-ID datasets, we assume all images of a person in one camera belong to a single tracklet. Then, we randomly assign a unique pseudo label to each tracklet for each camera.

4.2. Implement details

The structure of the backbone is shown in Fig. 2. To avoid overfitting and restrain negative pairs at the intra-camera learning stage, we add an embedding block at the top of the backbone, which contains two batch normalization layers and one drop-out layer. The training images are resized to 256 × 128. A camera evenly sampling strategy is used in training data loading, which randomly sample the same number images from each camera in a mini batch. The batch size of our experiments is set to 60. Adam optimizer is applied in our training process, with initializing the learning rate of $3.5e^{-4}$. The total training epoch is set to 80. The exemplar initiating epoch is set as 30 for MARS since it has most tracklets and is set as 10 for the other datasets. During the initializing epoch, we only compute intra-camera loss for initializing exemplars.

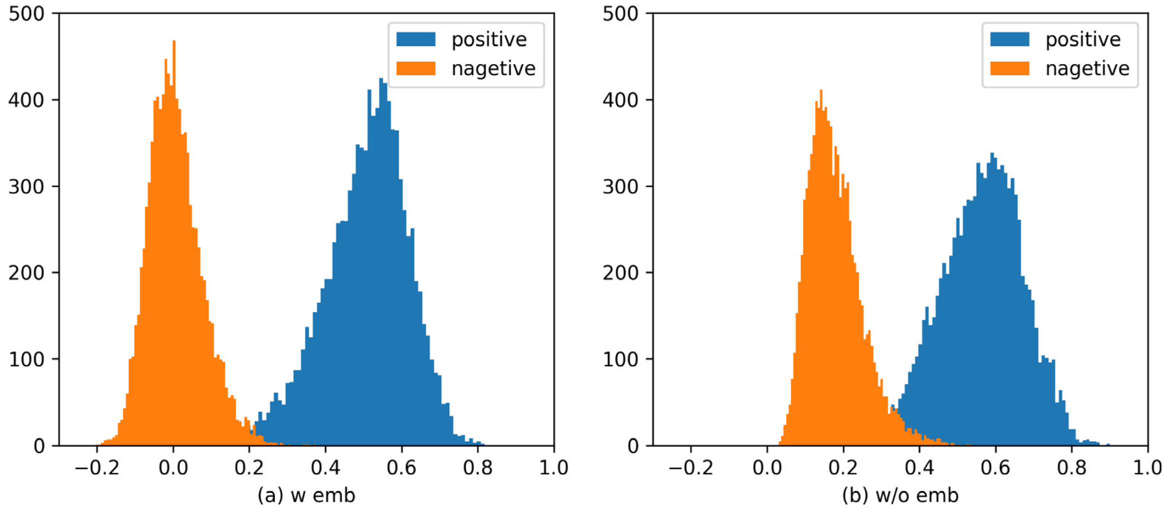
4.3. Ablation study

Structure & sampling strategy. We quickly verify the validity of the representation generator and camera evenly sampling strategy on MARS and Market-1501. Ablation studies are shown in Table 2. and Fig. 4. After adding a BN layer for both the supervised algorithm and unsupervised algorithm, the similarity score of negative samples becomes smaller than that of positive samples. Due to this, the positive and negative samples are much easier to be discriminated. As in previous work (Ioffe & Szegedy, 2015; Shen et al., 2018; Wu et al., 2019), BN of the

Table 3
Ablation of λ on image person RE-ID datasets.

Threshold	Market-1501		DukeMTMC-ReID		CUHK03		MSMT17	
Metric (%)	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
$\lambda = 0.55$	67.5	85.5	54.2	74.8	70.5	59.6	20.2	46.0
$\lambda = 0.60$	68.9	86.3	55.2	74.3	69.4	57.2	21.7	50.2
$\lambda = 0.65$	70.3	87.2	53.3	75.0	68.2	56.5	21.5	49.5
$\lambda = 0.70$	71.0	87.9	55.7	75.7	63.4	51.0	20.9	47.3
$\lambda = 0.75$	69.3	86.3	55.1	75.0	61.6	48.4	21.3	49.2
Average performance	69.4	86.6	54.3	75.0	66.6	54.5	21.2	48.4
Dynamic λ	66.6	85.4	55.4	75.6	73.1	62.5	20.1	45.5

1st and 2nd best results are in red/blue respectively.



(a) w emb
(b) w/o emb

Fig. 4. Similarity score distributions of positive pairs and negative pairs on Market-1501. (a) is the distribution of positive pairs; (b) is the distribution of negative pairs. After using the embedding block, positive pairs and negative pairs are easier to be distinguished.

Table 4
Ablation of λ on video person RE-ID datasets.

Metric (%)	MARS		Prid2011		iLIDS-VID	
	mAP	Rank-1	Rank-1	Rank-5	Rank-1	Rank-5
$\lambda = 0.55$	38.7	58.1	71.9	92.1	54.0	74.0
$\lambda = 0.60$	40.5	59.9	79.8	93.3	51.3	72.7
$\lambda = 0.65$	39.3	58.1	80.9	94.4	57.3	72.0
$\lambda = 0.70$	37.8	57.7	77.5	92.1	47.3	70.0
$\lambda = 0.75$	35.5	54.5	70.8	91.0	48.0	69.3
Average performance	38.4	57.7	76.2	92.6	51.6	71.6
Dynamic λ	44.8	61.5	82.0	96.6	60.0	82.7

1st and 2nd best results are in red/blue respectively.

embedding block may help the deep network converge faster and we find that the faster convergence helps better distinguish negative pairs. Considering that people's track always does not cover all of the cameras, different cameras have different number of tracklet. For example, in MARS, six cameras have 520, 447, 314, 195, 375 and 104 people respectively. We apply a camera evenly sampling strategy in data loading to balance the model learning speed over different cameras. The ablation study in Table 2, where KT denotes evenly K images from T cameras. Comparing with random sampling, camera evenly sampling averagely improves 12.3% Rank-1 and 10.5% mAP on these two datasets.

Threshold λ analysis. As shown in Table 3 and Table 4, we evaluate the performance of fixed λ from $\{0.55, 0.6, 0.65, 0.7, 0.75\}$ and dynamic selection threshold strategy. Results of the fixed λ indicate that the performance of different λ is quit different and different datasets have different suitable threshold. For example, on Prid2011, $\lambda = 0.65$ outperforms $\lambda = 0.75$ by 10.1% Rank-1. CUHK03 and iLIDS-VID perform better with small thresholds, while Market prefers large thresholds. To this end, we adapt dynamic threshold which dynamically changes on a wider range to select training pairs. As shown in Table 3, the dynamic λ has preponderant performance over the fixed λ on video person RE-ID. Although, the dynamic λ does not perform on all datasets, it averagely outperforms the fixed λ by 6.0% Rank-1 and 7.6% Rank-5. This makes us pay more attention on the unsupervised algorithm instead of hyper-parameter tuning.

Training time analysis. As shown in Table 7, we compare MARS training time on the 1080-Ti GPU to demonstrate the efficiency of EEA. UGA needs 12.6 GPU hours since wasting much time for two-stage training. For BUC, we run the code[2] released on the github. BUC takes much time for clustering, since it has to re-extract for all training data. It stores the feature of each cluster to compute repelled loss and the dimension of features is 2048. Hence, it takes 30.4 GPU hours (15.2 h and 2 1080-Ti GPUs) for training. TAUDL and UTAL do not release code. But comparing with EEA (60 batch size, Resnet-50 backbone, 80 epochs, 6.3 GPU hours), they set that the batch size is 384, the backbone is Resnet-50 and training epoch is 200. We estimate that they need 2.5 h and 5 1080-Ti GPUs for training at least, totally 12.5 GPU hours. Comparing with these methods, our end-to-end training methods EEA is much more efficient, only needing 6.3 GPU hours for training which is the half of UGA, TAUDL and UTAL. Comparing with

[2] https://github.com/vana77/Bottom-up-Clustering-Person-Re-identification.

Table 5
Ablation of intra-camera & inter-camera loss on image person RE-ID datasets.

Metric (%)	Market-1501		DukeMTMC-ReID		CUHK03		MSMT17	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
Intra-camera loss	54.8	77.5	52.5	72.6	56.3	42.2	19.7	45.7
Inter-camera loss	36.7	74.5	45.1	66.2	39.3	24.8	14.6	37.2
Total loss	66.6	85.4	55.4	75.6	73.1	62.5	20.1	45.5
Improvement	11.8	7.9	2.9	3.0	16.3	20.3	0.4	-0.2

Table 6
Ablation of intra-camera & inter-camera loss on video person RE-ID datasets.

Metric (%)	MARS		Prid2011		iLIDS-VID	
	mAP	Rank-1	Rank-1	Rank-5	Rank-1	Rank-5
Intra-camera loss	35.1	55.1	65.2	86.5	42.7	74.0
Inter-camera loss	35.6	52.7	70.8	91.0	50.7	78.7
Total loss	44.8	61.5	82.0	96.6	60.0	82.7
Improvement	9.7	6.4	16.8	10.1	17.3	8.7

Table 7
GPU hours of existing tracklet association methods on MARS.

Methods	Epochs	GPU occupying	GPU hours
TAUDL (Li, Zhu et al., 2018)	200	About 5	About 12.5
UTAL (Li et al., 2019)	200	About 5	About 12.5
BUC (Lin et al., 2019)	380	2	30.4
UGA (Wu et al., 2019)	160	1	12.6
EEA (This work)	**80**	**1**	**6.3**

existing methods in Tables 9 and 10, EEA outperforms most of them. The end-to-end framework EEA achieves a better trade-off between performance and training speed.

Loss functions analysis. In Table 5 and Table 6, we separately apply the intra-camera loss, the inter-camera loss and the total loss for training. The intra-camera loss achieves performs well by learning from single camera negative pairs. The inter-camera loss performs worse than the intra-camera loss. However, adding an inter-camera for view-invariant representations learning, Rank-1 averagely boosts 10.2% on these seven datasets. This demonstrates the importance of the intra-camera representation learning. Without it, inter-camera learning can be easily misled by noisy associations.

Noise analysis. The assumption of our experiments is one person has only one tracklet in each camera through SSTT sampling. However, it may not always hold in real-word applications. The ID duplication and mislabeling often occur in practice. The ID duplication is that the tracklets of the same person are given different pseudo labels. While the mislabeling is assigned the tracklets of different persons with the same pseudo labels. As shown in Table 8, we randomly select a part (10%, 20%, 50%) of persons per camera to simulate the ID duplication and mislabeling. We also count the standard deviation of these performances. According to Table 8, for ID duplication, Standard deviations of Rank-1 and mAP are 5.4 and 4.8, while they are 1.9 and 2.4 for mislabeling. ID duplication causes more damage. It should be avoided in real word applications.

Visual results. Several MARS examples of EEA retrieval results are shown in Fig. 5. Part of false matches come from the same camera, since the similar appearance (*i.e.*, the 2nd retrieval tracklet of the 3th raw and the 2nd retrieval tracklet of the 4th raw). The other false matches belong to the same camera (*i.e.*, the 1st, 2nd retrieval tracklets of the 2nd raw and the 1st, 3rd retrieval tracklets of the 5th raw). One possible reason is that the false pedestrian detection introduces a lot of background in pedestrian bounding box (Chi et al., 2019a, 2019b; Zhang et al., 2018). In these false matches, background is the same and accounts for a large proportion. The unsupervised trained model regards these false matches as the same pedestrian. This issue demonstrates that the cross camera constrain is necessary for KNN graph building (Eq. (6)). Without this constrain, KNN graph would link false matches belonging to the same camera. It is invalid for inter-camera representation learning and harmful for intra-camera representation learning.

4.4. Comparison to the state-of-the-art methods

We compare our EEA with some state-of-the-art unsupervised person RE-ID methods, specifically comparing with four similar tracklet association methods. The performances of these methods are shown in Tables 9 and 10. We also report the average performance (mAP, Rank-1, Rank-5 and Rank-20) and corresponding standard deviations of 3 runs in Table 11.

Image person RE-ID datasets. Table 9 shows the performance of several state-of-the-art methods on four image person RE-ID datasets, containing four GAN based methods (HHL, SPGAN, SPGAN+LMP), two domain adaptation methods (TJ-AIDL, ECN), four unsupervised clustering methods (BUC, CAMEL, PUL and CDS) and two tracklet based method (UTAL, TAUDL). Comparing with them, UGA and EEA have the Overwhelming performance. UGA averagely outperforms the second by 9.6% on Rank-1 accuracy and 16.8% on mAP, respectively. EEA boosts 0.6% Rank-1 and 6% Rank-1 on DukeMTMC-ReID and CUHK03 respectively, achieving the best performance on DukeMTMC-ReID and CUHK03. Specifically, EEA cuts the training (UGA, UTAL and TAUDL) time in half.

Video person RE-ID datasets. We compare the proposed EEA on three video person RE-ID datasets with several state-of-the-art approaches in Table 10. EEA nearly achieves the best performance on both of three video person RE-ID datasets. It outperforms the second 1.1% Rank-1 and 2.7% Rank-1 on PRID2011 and iLIDS-VID respectively. Although EUG outperforms it 1.2% Rank-1 on MARS, EUG is a one-shot learning algorithm and EEA is a totally unsupervised algorithm. Furthermore, EUG is sensitive to the enlarging factors and UGA is a little sensitive to the threshold λ. On MARS, EUG declines from 62.67% to 42.77% with the enlarging factors changing, and UGA declines from 59.9% to 54.5% with the λ changing. EEA adapts a dynamic threshold strategy to alleviate this problem. Comparatively, EEA is more robust to the hyper-parameter. Comparing with Snippet and IANet, EEA largely closes the gap between supervised learning and unsupervised learning on the person RE-ID task.

Comparison with the unsupervised graph based methods. We compare our EEA with the existed graph based work (*i.e.*, TUADL (Li, Zhu et al., 2018), UTAL (Li et al., 2019), RACE (Ye et al., 2018) and ECN (Zhong, Zheng, Luo et al., 2019)) in Tables 9 and 10. EEA averagely outperforms TAUDL by (17.6% on Rank-1) in image person RE-ID datasets and (27.9% on Rank-1) in video person RE-ID datasets. EEA outperforms UTAL by (12.3% on Rank-1,) in image person RE-ID datasets and (12.5% on Rank-1) in video person RE-ID datasets. In addition, both of TAUDL and UTAL match the positive pairs in the mini batch which needs a large batch size (384) to sample the underlying positive pairs and may occupy at least five 1080Ti GPUs in training. But EEA can be implemented on one 1080-Ti, since the exemplar memory

Table 8
Analysis of noisy tracklets on MARS.

Noise	ID duplication				Mislabeling			
Metric (%)	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10
10%	35.5	53.1	68.2	73.5	43.3	60.2	74.2	78.4
20%	34.0	52.7	65.9	72.4	41.6	59.9	72.9	78.2
50%	24.5	41.4	57.5	63.4	37.5	56.0	70.4	74.7
average	31.3	49.1	63.9	69.8	**40.8**	**58.7**	**72.5**	**77.1**
std	4.8	5.4	4.5	4.5	**2.4**	**1.9**	**1.6**	**1.7**

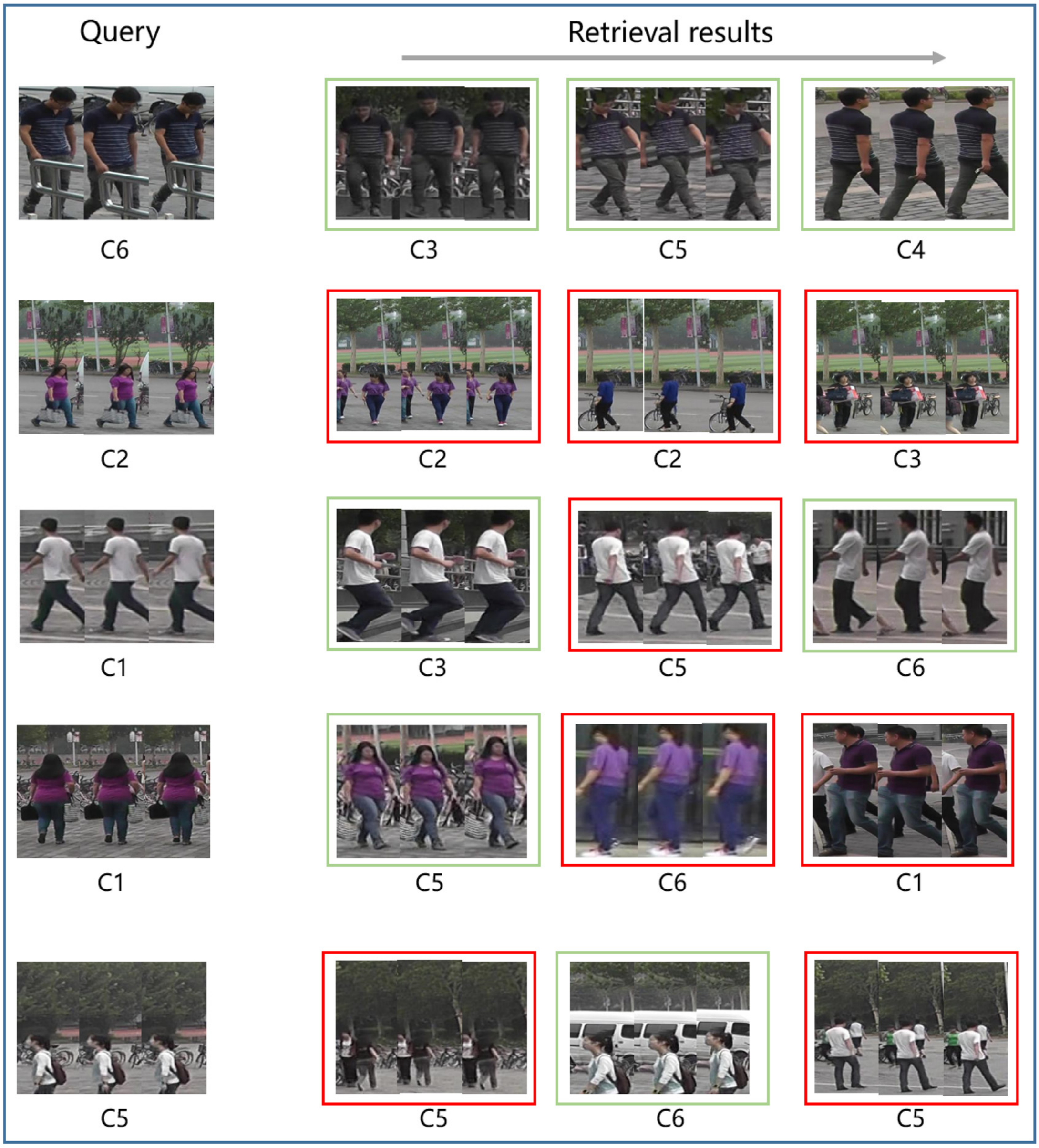


Fig. 5. Example tracklets retrieved by EEA model among unlabeled short fragmented tracklets. Each row denotes a case. The first tracklet is given as a query, while the remaining three are retrieval results. The green and red bounding box denote the true/false retrieval results, respectively. C1, C2, ..., C6 denote the 1st, 2nd, . . . , 6th camera, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

module allows using a small mini batch in training. Different from RACE (Ye et al., 2018) merging the underlying positive tracklets directly, EEA uses the cross-camera loss and cross-camera graph to associate tracklets. It is more robust to noisy associations. Due to this, EEA easily achieves the higher performance than RACE. Comparing with ECN, EEA averagely outperforms

Table 9
Comparing UGA with the state-of-the-art methods on the image person RE-ID dataset.

Datasets	Reference	Market-1501		DukeMTMC-ReID		CUHK03		MSMT17	
Metric (%)		mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
DECAMEL (Yu, Wu, & Zheng, 2018)	TPAMI'19	32.4	60.2	–	–	–	38.27	11.1	30.3
LOMO (Liao, Hu, Zhu, & Li, 2015)	CVPR'15	8.0	27.2	4.8	12.3	–	46.25	–	–
BoW (Zheng et al., 2015)	ICCV'15	14.8	35.8	8.3	17.1	–	–	–	–
DIC (Kodirov, Xiang, & Gong, 2015)	BMVC'15	22.7	50.2	–	–	–	36.5	–	–
UDML (Peng, Xiang, Wang, Pontil, Gong, Huang, & Tian, 2016)	CVPR'16	12.4	34.5	7.3	12.3	–	–	–	–
HHL (Zhong, Zheng, Li et al., 2018)	ECCV'18	31.4	62.2	27.2	46.9	–	–	–	–
SPGAN (Deng et al., 2018)	CVPR'18	22.8	51.5	22.3	41.1				
SPGAN+LMP (Deng et al., 2018)	CVPR'18	26.7	57.7	26.2	46.4	–	–	–	–
TJ-AIDL (Wang et al., 2018)	CVPR'17	26.5	58.2	23.0	44.3	–	–	–	–
PTGAN (Wei et al., 2018)	CVPR'18	15.7	38.6	13.5	27.4	–	37.5	–	–
CAMEL (Yu, Wu, & Zheng, 2017)	ICCV'17	26.3	54.5	–	–	–	39.4	–	–
PUL (Fan et al., 2018)	ToMM'18	20.1	44.7	16.4	30.4	–	–	–	–
BUC (Lin et al., 2019)	AAA'19	38.3	66.2	27.5	47.4	–	–	–	–
MAR (Yu et al., 2019)	CVPR'19	40.0	67.7	48.0	67.1	–	–	–	–
CDS (Jinlin et al., 2018)	ICME'19	39.9	71.6	42.7	67.2	–	–	–	–
ECN (Zhong, Zheng, Luo et al., 2019)	CVPR'19	43.0	75.1	40.4	63.3	–	–	10.2	30.2
TAUDL (Li, Zhu et al., 2018)	ECCV'18	41.2	63.7	43.5	61.7	31.2	44.7	12.5	28.4
UTAL (Li et al., 2019)	TPAMI'19	46.2	69.2	44.6	62.3	42.3	56.3	13.1	31.4
UGA (Wu et al., 2019)	ICCV'19	70.3	87.2	53.3	75.0	68.2	56.5	21.7	49.5
EEA	This work	66.6	85.4	55.4	75.6	73.1	62.5	20.1	45.5
PCB (Sun et al., 2018)[a]	EECV'18	77.4	92.3	69.3	83.3	–	–	–	–
GCS (Chen, Xu et al., 2018)[a]	CVPR'18	81.6	93.5	69.5	84.9	97.2	88.8	–	–
SFT (Luo, Chen, Wang, & Zhang, 2019)[a]	ICCV'19	82.7	93.4	73.2	86.9	–	–	47.6	73.6

1st, 2nd, 3rd best results are in red/blue/green respectively.

[a] Denotes the supervised algorithm.

Table 10
Comparing UGA with the state-of-the-art methods on the video person RE-ID dataset.

Datasets	PRID2011			iLIDS-VID			MARS		
Metric (%)	R1	R5	R20	R1	R5	R20	R1	R5	mAP
SMP (Liu, Wang, & Lu, 2017)	80.9	95.6	99.4	41.7	66.3	80.7	23.9	35.8	10.5
DGM+MLAPG (Ye, Ma, Zheng, Li, & Yuen, 2017)	73.5	92.6	99.0	37.1	61.3	82.0	24.6	42.6	11.8
DGM+IDE (Ye et al., 2017)	56.4	81.3	96.4	36.2	62.8	82.7	36.8	54.0	21.3
DASy (Bak et al., 2018)	43.0	–	–	56.5	–	–	–	–	–
GRDL (Kodirov, Xiang, Fu, & Gong, 2016)	41.6	76.4	89.9	25.7	49.9	77.6	19.3	33.2	9.56
DTW (Ma et al., 2016)	41.7	67.1	90.1	31.5	62.1	82.4	–	–	–
BUC (Lin et al., 2019)	–	–	–	–	–	–	61.1	75.1	38.0
EUG (Wu et al., 2018)[b]	–	–	–	–	–	–	62.7	74.9	42.5
RACE (Ye et al., 2018)	50.6	79.4	91.8	19.3	39.3	68.7	43.2	57.1	24.5
TAUDL (Li, Zhu et al., 2018)	49.4	78.7	98.9	26.7	51.3	82.0	43.8	59.9	29.1
UTAL (Li et al., 2019)	54.7	83.1	96.2	35.1	59.0	83.8	49.9	66.4	35.2
UGA (Wu et al., 2019)	80.9	94.4	100	57.3	72.0	87.3	58.1	73.4	39.3
EEA (This work)	82.0	96.6	100	60.0	82.7	94.0	61.5	76.5	44.8
Snippet (Chen, Li, Xiao, Yi and Wang, 2018)[a]	93.0	99.3	100.0	85.4	96.7	99.5	86.3	94.7	76.1
IANet (Hou et al., 2019a)[a]	–	–	–	54.6	79.4	86.9	84.0	93.7	73.3

1st and 2nd best results are in red/blue respectively.

R1, R5 and R20 denote Rank-1, Rank-5 and Rank-20 respectively.

[a] Denotes the supervised algorithm.

[b] EUG reports results with the hyper-parameter $p = 0.5$.

Table 11
Average results of three runs.

Metric (%)	mAP	Rank-1	Rank-5	Rank-20
Market1501	68.7 ± 0.309	86.2 ± 0.779	94.2 ± 0.161	97.87 ± 0.125
DukeMTMC	54.2 ± 0.838	73.6 ± 0.533	84.1 ± 0.850	90.6 ± 0.535
CUHK03	72.5 ± 0.816	61.3 ± 0.816	86.4 ± 0.873	95.8 ± 0.356
MSMT17	20.5 ± 0.694	46.4 ± 1.275	60.7 ± 1.241	73.1 ± 1.080
MARS	–	44.6 ± 0.939	75.8 ± 0.408	84.1 ± 0.262
Prid2011	–	81.6 ± 0.519	81.6 ± 0.519	99.9 ± 0.125
iLIDS-VID	–	60.9 ± 1.657	82.9 ± 0.829	95.3 ± 0.531

ECN by (9.5% on Rank-1, 12.1% on mAP) in image person RE-ID datasets. Because ECN simply apply a KNN graph to associate the underlying positive samples, while EEA uses the more precise graph (cross-camera graph) to associate the underlying positive pairs.

5. Conclusion

In this paper, we have proposed a novel yet effective End-to-End Exemplar Association (EEA) approach to address the unsupervised person RE-ID problem. In this work, we investigate unsupervised person RE-ID representation from inter-camera representation learning and intra-camera representation learning. We develop an exemplar memory module to simultaneously learn both of them. Based on this, we propose three strategies: (1) **end-to-end exemplar-based training**, (2) **exemplar association** and (3) **dynamic selection threshold**. The first one is to accelerate the training process, while the others aim to improve the tracklet association correctness. Due to this, EEA achieves competitive performances and cuts the training time of existing tracklet association methods (UTAL, TAUDL, UGA) in a half. Experiments on four image RE-ID datasets and three video RE-ID datasets demonstrate the superiority of EEA.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the National Key Research and Development Plan under Grant 2019YFC2003901, and the Chinese National Natural Science Foundation Projects #61806203, #61876178, #61872367, #61976229.

References

Bai, S., Bai, X., & Tian, Q. (2017). Scalable person re-identification on supervised smoothed manifold. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2530–2539).

Bak, S., Carr, P., & Lalonde, J.-F. (2018). Domain adaptation through synthesis for unsupervised person re-identification. arXiv preprint arXiv:1804.10094.

Bourlard, H., & Kamp, Y. (1988). Auto-association by multilayer perceptrons and singular value decomposition. *Biological Cybernetics*, *59*(4–5), 291–294.

Chang, X., Hospedales, T. M., & Xiang, T. (2018). Multi-level factorisation net for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2109–2118).

Chen, D., Li, H., Xiao, T., Yi, S., & Wang, X. (2018). 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).

Chen, D., Xu, D., Li, H., Sebe, N., & Wang, X. (2018). Group consistent similarity learning via deep crf for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8649–8658).

Chi, C., Zhang, S., Xing, J., Lei, Z., Li, S. Z., & Zou, X. (2019a). Pedhunter: Occlusion robust pedestrian detector in crowded scenes. arXiv preprint arXiv:1909.06826.

Chi, C., Zhang, S., Xing, J., Lei, Z., Li, S. Z., & Zou, X. (2019b). Relational learning for joint head and human detection. arXiv preprint arXiv:1909.10674.

Deng, W., Zheng, L., Ye, Q., Kang, G., Yang, Y., & Jiao, J. (2018). Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 994–1003).

Dollar, P., Wojek, C., Schiele, B., & Perona, P. (2012). Pedestrian detection: An evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(4), 743–761.

Dosovitskiy, A., Fischer, P., Springenberg, J. T., Riedmiller, M., & Brox, T. (2015). Discriminative unsupervised feature learning with exemplar convolutional neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(9), 1734–1747.

Fan, H., Zheng, L., Yan, C., & Yang, Y. (2018). Unsupervised person re-identification: Clustering and fine-tuning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, *14*(4), 83.

Gong, C., Liu, T., Tao, D., Fu, K., Tu, E., & Yang, J. (2015). Deformed graph laplacian for semisupervised learning. *IEEE Transactions on Neural Networks and Learning Systems*, *26*(10), 2261–2274.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. In *Advances in neural information processing systems* (pp. 2672–2680).

Hadsell, R., Chopra, S., & LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06), Vol. 2* (pp. 1735–1742). IEEE.

He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2019). Momentum contrast for unsupervised visual representation learning. arXiv preprint arXiv:1911.05722.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, *313*(5786), 504–507.

Hirzer, M., Beleznai, C., Roth, P. M., & Bischof, H. (2011). Person re-identification by descriptive and discriminative classification. In *Scandinavian conference on image analysis* (pp. 91–102). Springer.

Hou, R., Ma, B., Chang, H., Gu, X., Shan, S., & Chen, X. (2019a). Interaction-and-aggregation network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 9317–9326).

Hou, R., Ma, B., Chang, H., Gu, X., Shan, S., & Chen, X. (2019b). Vrstc: Occlusion-free video person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7183–7192).

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167.

Jinlin, W., Shengcai, L., Zhen, L., Xiaobo, W., Yang, Y., & Li, S. Z. (2018). Clustering and dynamic sampling for unsupervised domain adaptation in person re-identification. In *IEEE international conference on multimedia and expo (ICME)*.

Kodirov, E., Xiang, T., Fu, Z., & Gong, S. (2016). Person re-identification by unsupervised l_1 graph learning. In *European conference on computer vision* (pp. 178–195). Springer.

Kodirov, E., Xiang, T., & Gong, S. (2015). Dictionary learning with iterative laplacian regularisation for unsupervised person re-identification.. In *BMVC, Vol. 3* (p. 8).

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Lealtaixe, L., Milan, A., Reid, I. D., Roth, S., & Schindler, K. (2015). Motchallenge 2015: Towards a benchmark for multi-target tracking. *arXiv: Computer Vision and Pattern Recognition*.

Li, Y.-J., Yang, F.-E., Liu, Y.-C., Yeh, Y.-Y., Du, X., & Frank Wang, Y.-C. (2018). Adaptation and re-identification network: An unsupervised deep transfer learning approach to person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 172–178).

Li, W., Zhao, R., Xiao, T., & Wang, X. (2014a). Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 152–159).

Li, W., Zhao, R., Xiao, T., & Wang, X. (2014b). DeepReID: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Li, M., Zhu, X., & Gong, S. (2018). Unsupervised person re-identification by deep learning tracklet association. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 737–753).

Li, M., Zhu, X., & Gong, S. (2019). Unsupervised tracklet person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1.

Liang, W., Wang, G., Lai, J., & Zhu, J. (2018). M2m-gan: Many-to-many generative adversarial transfer learning for person re-identification. arXiv preprint arXiv:1811.03768.

Liao, S., Hu, Y., Zhu, X., & Li, S. Z. (2015). Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2197–2206).

Lin, Y., Dong, X., Zheng, L., Yan, Y., & Yang, Y. (2019). A bottom-up clustering approach to unsupervised person re-identification. In *AAAI conference on artificial intelligence, Vol. 2*.

Liu, Z., Wang, D., & Lu, H. (2017). Stepwise metric promotion for unsupervised video person re-identification. In *Proceedings of the IEEE international conference on computer vision* (pp. 2429–2438).

Liu, H., Zhu, X., Lei, Z., & Li, S. Z. (2019). AdaptiveFace: Adaptive margin and sampling for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 11947–11956).

Luo, C., Chen, Y., Wang, N., & Zhang, Z. (2019). Spectral feature transformation for person re-identification. In *Proceedings of the IEEE international conference on computer vision* (pp. 4976–4985).

Luo, Y., Zhu, J., Li, M., Ren, Y., & Zhang, B. (2018). Smooth neighbors on teacher graphs for semi-supervised learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8896–8905).

Ma, X., Zhu, X., Gong, S., Xie, X., Hu, J., Lam, K. M., et al. (2016). Person re-identification by unsupervised video matching. *Pattern Recognition*, *65*(C), 197–210.

Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., & Frey, B. (2015). Adversarial autoencoders. arXiv preprint arXiv:1511.05644.

Peng, P., Xiang, T., Wang, Y., Pontil, M., Gong, S., Huang, T., & Tian, Y. (2016). Unsupervised cross-dataset transfer learning for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1306–1315).

Ristani, E., Solera, F., Zou, R., Cucchiara, R., & Tomasi, C. (2016a). Performance measures and a data set for multi-target, multi-camera tracking. In *European conference on computer vision* (pp. 17–35). Springer.

Ristani, E., Solera, F., Zou, R. S., Cucchiara, R., & Tomasi, C. (2016b). Performance measures and a data set for multi-target, multi-camera tracking. In *European conference on computer vision* (pp. 17–35).

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85–117.

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 815–823).

Shen, Y., Li, H., Yi, S., Chen, D., & Wang, X. (2018). Person re-identification with deep similarity-guided graph neural network. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 486–504).

Shi, H., Yang, Y., Zhu, X., Liao, S., Lei, Z., Zheng, W., et al. (2016). Embedding deep metric for person re-identification: A study against large variations. In *European conference on computer vision* (pp. 732–748). Springer.

Sun, Y., Zheng, L., Deng, W., & Wang, S. (2017). Svdnet for pedestrian retrieval. *ICCV*, 3820–3828.

Sun, Y., Zheng, L., Yang, Y., Tian, Q., & Wang, S. (2018). Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European conference on computer vision (ECCV)* (pp. 480–496).

Tan, Z., Yang, Y., Wan, J., Hang, H., Guo, G., & Li, S. Z. (2019). Attention based pedestrian attribute analysis. *IEEE Transactions on Image Processing*.

Tavanaei, A., Ghodrati, M., Kheradpisheh, S. R., Masquelier, T., & Maida, A. (2018). Deep learning in spiking neural networks. *Neural Networks*.

Wang, T., Gong, S., Zhu, X., & Wang, S. (2014). Person re-identification by video ranking. In *European conference on computer vision* (pp. 688–703). Springer.

Wang, G., Lai, J., Huang, P., & Xie, X. (2019). Spatial-temporal person re-identification. In *National conference on artificial intelligence*.

Wang, G., Yang, Y., Cheng, J., Wang, J., & Hou, Z. (2019). Color-sensitive person re-identification. In *IJCAI 2019: 28th international joint conference on artificial intelligence*.

Wang, J., Zhu, X., Gong, S., & Li, W. (2018). Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2275–2284).

Wei, L., Zhang, S., Gao, W., & Tian, Q. (2018). Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 79–88).

Wu, Y., Lin, Y., Dong, X., Yan, Y., Ouyang, W., & Yang, Y. (2018). 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).

Wu, J., Yang, Y., Liu, H., Liao, S., Lei, Z., & Li, S. Z. (2019). Unsupervised graph association for person re-identification. In *Proceedings of the IEEE international conference on computer vision* (pp. 8321–8330).

Ye, M., Lan, X., & Yuen, P. C. (2018). Robust anchor embedding for unsupervised video person re-identification in the wild. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 170–186).

Ye, M., Ma, A. J., Zheng, L., Li, J., & Yuen, P. C. (2017). Dynamic label graph matching for unsupervised video re-identification. In *Proceedings of the IEEE international conference on computer vision* (pp. 5142–5150).

Yi, D., Lei, Z., Liao, S., & Li, S. Z. (2014). Deep metric learning for person re-identification. In *2014 22nd international conference on pattern recognition* (pp. 34–39). IEEE.

Yu, H.-X., Wu, A., & Zheng, W.-S. (2017). Cross-view asymmetric metric learning for unsupervised person re-identification. In *Proceedings of the IEEE international conference on computer vision* (pp. 994–1002).

Yu, H.-X., Wu, A., & Zheng, W.-S. (2018). Unsupervised person re-identification by deep asymmetric metric embedding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yu, H., Zheng, W., Wu, A., Guo, X., Gong, S., & Lai, J. (2019). Unsupervised person re-identification by soft multilabel learning. (pp. 2148–2157).

Zhang, S., Benenson, R., & Schiele, B. (2017). Citypersons: A diverse dataset for pedestrian detection. (pp. 4457–4465).

Zhang, S., Wen, L., Bian, X., Lei, Z., & Li, S. Z. (2018). Occlusion-aware r-cnn: Detecting pedestrians in a crowd. (pp. 637–653).

Zheng, L., Bie, Z., Sun, Y., Wang, J., Su, C., Wang, S., et al. (2016). Mars: A video benchmark for large-scale person re-identification. In *European conference on computer vision* (pp. 868–884). Springer.

Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., & Tian, Q. (2015). Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision* (pp. 1116–1124).

Zheng, Z., Zheng, L., & Yang, Y. (2017). Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *International conference on computer vision* (pp. 3774–3782).

Zheng, Z., Zheng, L., & Yang, Y. (2018). A discriminatively learned cnn embedding for person reidentification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, *14*(1), 13.

Zhong, Z., Zheng, L., Li, S., & Yang, Y. (2018). Generalizing a person retrieval model hetero-and homogeneously. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 172–188).

Zhong, Z., Zheng, L., Luo, Z., Li, S., & Yang, Y. (2019). Invariance matters: Exemplar memory for domain adaptive person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 598–607).

Zhong, Z., Zheng, L., Zheng, Z., Li, S., & Yang, Y. (2018). Camstyle: A novel data augmentation method for person re-identification. *IEEE Transactions on Image Processing*, *28*(3), 1176–1190.

Zhong, Z., Zheng, L., Zheng, Z., Li, S., & Yang, Y. (2019). Camstyle: A novel data augmentation method for person re-identification. *IEEE Transactions on Image Processing*, *28*(3), 1176–1190.

Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 2223–2232).