Unsupervised Graph Association for Person Re-identification

Jinlin Wu*[1,2], Yang Yang*[1,2], Hao Liu[1,2], Shengcai Liao[3], Zhen Lei †[1,2], and Stan Z. Li[1,2]

[1]CBSR & NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing, China.
[2]University of Chinese Academy of Sciences, Beijing, China.
[3]Inception Institute of Artificial Intelligence (IIAI), Abu Dhabi, UAE,
{ jinlin.wu, yang.yang, hao.liu2016, zlei, szli}@nlpr.ia.ac.cn, scliao@ieee.org

Abstract

In this paper, we propose a novel unsupervised graph association (UGA) to learn the underlying view-invariant representations from the video pedestrian tracklets. The core points of it are mining the cross-view relationships and reducing the damage of noisy associations. To this end, UGA adopts a two-stage training strategy: (1) ***intra-camera learning stage*** *and (2)* ***inter-camera learning stage****. The former is to learn representations of a person with regards to camera information, which helps to reduce false cross-view associations in the second stage. Compared with existing tracklet-based methods, ours can build more accurate cross-view associations and require lower GPU memory. Extensive experiments and ablation studies on seven RE-ID datasets demonstrate the superiority of the proposed UGA over most state-of-the-art unsupervised and domain adaptation RE-ID methods. Code is available at github*[1].

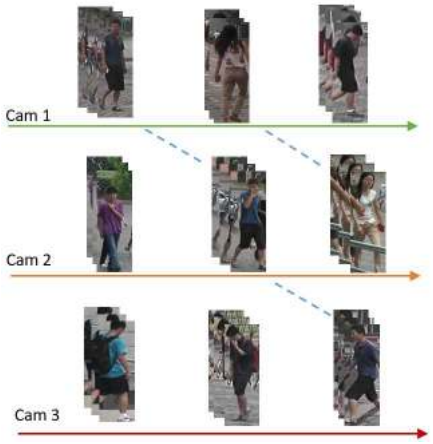

Figure 1: Examples of the pair-wise labelled tracklets. Pair-wise labelled tracklets refer the images belonging to the same person under different cameras.

1. Introduction

Person Re-identification (RE-ID) aims to match the same pedestrian across non-overlapping camera views, which has potential applications like longterm multi-camera tracking and forensic search. Benefiting from the advance of deep learning, especially the deep convolution network [14, 10], the performance of RE-ID has obtained significant improvements [19, 42, 30, 32, 29, 31, 48]. In supervised learning, the deep CNN learns view-invariant representations from the pair-wise labelled data. Since deep CNN is a data-driven method, it requires a large number of pair-wise labelled data in training. Figure 1 show some pair-wise labelled tracklets from different cameras. However, labelling sufficient pair-wise RE-ID data is expensive and time-consuming. How to improve the performance and scalability of deep RE-ID algorithm without pair-wise labelled data (*i.e.,* unsupervised learning) is a great challenge in recent person RE-ID research.

There have been a series of unsupervised image based methods to address this problem, which can be roughly divided into three categories: 1) image-to-image translation, 2) domain adaptation, 3) unsupervised clustering. The image-to-image translation methods [51, 49, 4, 1, 38] transfer the source domain images to the target domain by GAN [9] network. The domain adaptation methods [21, 36] aim to transfer the source domain trained model to the target domain in an unsupervised manner. Unsupervised clustering methods [5, 16] obtain the pseudo labels of target domain data through the unsupervised clustering algorithms

*Equal Contribution
†Corresponding author
[1]https://github.com/yichuan9527/Unsupervised-Graph-Association-for-Person-Re-identification

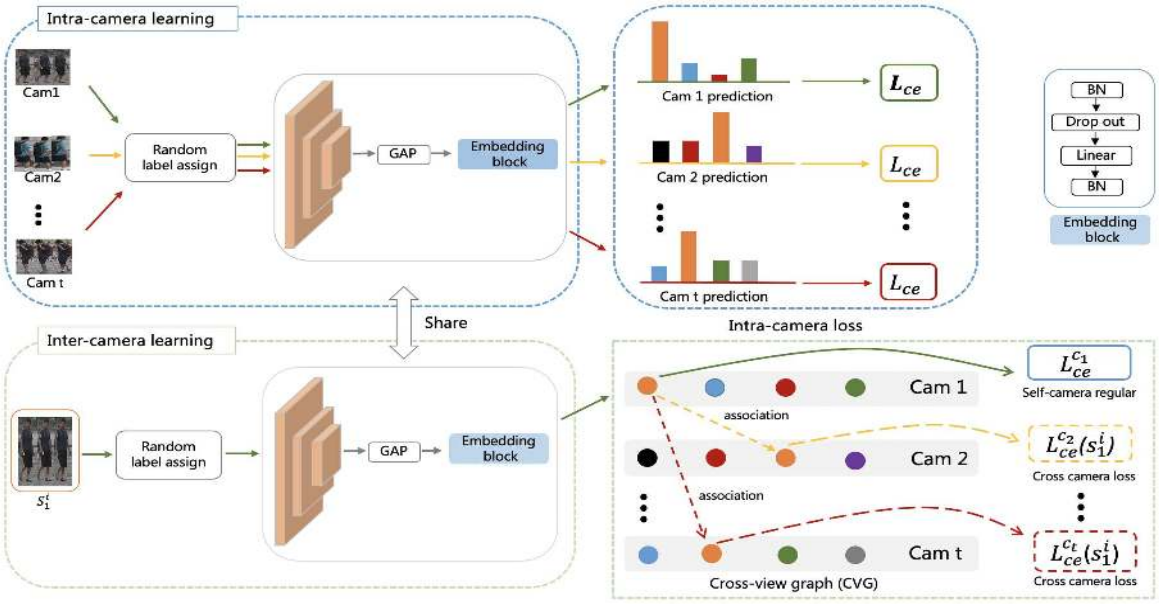

Figure 2: The framework of the proposed unsupervised graph association (UGA), including 1) intra-camera learning stage and 2) inter-camera learning stage. The model architecture consists of a Resnet-50 backbone, a global average pooling layer (GAP), an embedding block and multi-branch-classifier. The embedding block includes a batch normalization layer, a drop-out layer, a FC layer reducing the 2048-dim feature to 1024-dim and a batch normalization layer.

and fine tune the source domain model with pseudo labels on target domain.

However, the precondition of above mentioned methods is that there are some similarities between the source domain and the target domain. For example, as shown in [5, 4, 22], the above mentioned three methods can easily achieve high performances between Market1501 [46] and DukeMTMC-ReID [27], since Market1501 and DukeMTMC-ReID are similar to each other. However, the performance becomes worse when using MSMT17 [38] as the target. This is because the variations of illumination and resolution are more complicated in MSMT17 than that in Market-1501 and DukeMTMC-ReID. The unsupervised image based methods are sensitive to these variations and have poor scalability to unknown scenes.

Recently, the tracklet based methods (*i.e.*, TAUDL [17], UTAL [18], RACE [40], BUC [23]) have been proposed to overcome this weakness. Pedestrian tracklets are easily obtained by existing tracking algorithms [15, 28, 6, 7, 44]. The frames of the same tracklet generally belong to the same identity. Owing to this, the complexity of the unsupervised learning is reduced as presented in TAUDL, UTAL. However, UTAL and TAUDL match the underlying positive pairs in the mini batch. Due to this, both of them need a large batch size (384) to sample the underlying positive pairs which may occupy at least five 1080-Ti GPUs in training. RACE and BUC, which progressively merge the underlying positive pairs in training, are easily damaged by merging noisy pairs.

To address these problems, we propose an unsupervised graph association (UGA) framework for tracklet based unsupervised RE-ID. The pip line is shown in Figure 2, it contains an **intra-camera learning stage** and an **inter-camera learning stage**.

Intra-camera learning stage. We apply the multiple-branch-classifier (MBC) structure to learn the intra-camera representation, where each classifier branch corresponds to one camera's classification task. Besides, we apply an embedding block at the top of the backbone, which makes the negative pairs easier to be distinguished and avoids the training overfitting.

Inter-camera learning stage. We build a cross-view graph (CVG) to associate pedestrian tracklets and develop a cross-camera loss to learn the view-invariant representations from CVG. We replace the weights of MBC with the corresponding nodes of CVG to fast updating CVG in the training process. In order to reduce the damage of the noisy associations, we introduce two constraints (threshold, symmetry) into CVG and use the CVG's edge weight as the weighting of the cross-camera loss. To sum up, the contributions of this paper can be summarized as follows:

- We propose a simple yet effective unsupervised person RE-ID framework, named unsupervised graph association (UGA). Without any source domain pre-training, UGA achieves high performance, with low GPU mem-

ory occupation.

- We incorporate a novel cross-view graph (CVG) and a cross-camera loss into UGA framework. By using both of them, model can learn the view-invariant representations from the underlying positive samples.
- We conduct extensive experiments and ablation studies on seven RE-ID datasets to demonstrate the effectiveness of the proposed UGA.

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 algorithm, person RE-ID has achieved a remarkable progress [42, 32, 34, 29, 2, 48, 35, 33]. Yi *et al.* [42] adopt image pairs and introduce part priors into a siamese network for learning the view-invariant representations. Sun *et al.* [32] and chang *et al.* [2] develop the part feature based methods to enhance the discriminative of Re-ID features. Wang *et al.* [34] 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 labelled data for training. To solve this problem, lots of unsupervised person RE-ID methods are proposed [51, 49, 21, 36, 5, 52, 47]. Zhong *et al.* [51, 49], Deng *et al.* [4] and Bake *et al.* [1] 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 *et al.* [21] and Wang *et al.* [37] apply the domain adaptation methods transferring source domain knowledge to target domain. Fan *et al.* [5] and Wu *et al.* [16] 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 *et al.* [17, 18] match the underlying positive pairs in the mini batch, using a cross camera histogram loss to learn the view-invariant features. Ye *et al.* [40] propose a robust embedding to reduce the damage of the noisy frames for estimator pseudo labels more accuracy.

Graph based methods. Considering the relationships between the training samples, graph based methods [25, 8, 3, 29] are used to provide more supervision signals for both of semi-supervised learning and supervised training. Luo *et al.* [25] 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 labelled 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.* [29] 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 labelled samples to build the graph for training.

3. Method

Definition. Suppose we have a dataset, captured from T cameras. We adopt the sparse space-time tracklets sampling (SSTT) [17] to sample the training tracklets $\{s_t^i, y_t^i\}$ from each camera. Denoting $s_t^i = \{I_1^{s_t^i}, I_2^{s_t^i}, ..., I_n^{s_t^i}\}$, where $I_n^{s_t^i}$ is the n-th image of the i-th tracklet ($i \in [1, \ldots, M_t]$) in t-th camera ($t \in [1, \ldots, T]$). We randomly assign a unique pseudo label $y_t^i (y_t^i \in \{y_t^1, \ldots, y_t^{M_t}\})$ for the s_t^i. $\phi(\cdot)$ is the backbone function. f_t is the t-th branch classifier of MBC. W_t^i is the weight of f_t, corresponding to the class of s_t^i.

3.1. Intra-camera learning

Through the SSTT sampling, we can obtain the training data $\{s_t^i, y_t^i\}$ for each camera and a person has at most one tracklet in each camera. To avoid the conflict of pseudo labels, we adopt the multi-task training to learn the intra-camera representation. The pip line of intra-camera learning is shown in Figure 2. A multi-branch-classifier (MBC) structure is adopted to model the persons classification in different cameras as a multi-task problem. All of the classifiers share the backbone features. For the $t-th$ branch, the softmax cross-entropy loss function is formulated as follows:

$$l_{ce}^t(I_n^{s_t^i}) = -\sum_{j=1}^{M_t} log(\frac{e^{(W_t^j)^T \phi(I_n^{s_t^i})}}{\sum_{k=1}^{M_t} e^{(W_t^k)^T \phi(I_n^{s_t^i})}}) \quad (1)$$

In the Eq. (1), W_t^j is the weight of s_t^i and $\phi(I_n^{s_t^i})$ is the representation vector of $I_n^{s_t^i}$ extracted by the backbone $\phi(\cdot)$. The total loss of all branches l_{intra} can be defined as Eq. (2), where N_{bs} denotes the batch size.

$$l_{intra} = \frac{1}{N_{bs}} \sum_{N_{bs}} l_{ce}^t(I_n^{s_t^i}) \quad (2)$$

To avoid overfitting and restrain negative pairs at the intra-camera learning stage, we add an embedding block at the top of the backbone (shown in Figure 2), which contains two batch normalization layers and one drop-out layer. As shown in Figure 3, the batch normalization layer is effective to reduce the average similarity score of the negative

Algorithm 1: Unsupervised graph association (UGA)

Input: Pair-wise unlabelled tracklets s_t^i of T cameras.
The Backbone ϕ and multi-branch-classifier f_t.
W_t^i is the weight vector of f_t.
Threshold λ and max iteration ep_{max}.
Initializing iteration step $ep \leftarrow 0$.
$(t = 1, \ldots, T)$

while $ep < ep_{max}$ **do**
1: $ep \leftarrow ep + 1$;
2: Computing L_{intra} according Eq. (2);
3: Updating ϕ and f_t;
end
1: Computing tracklets center c_t^i ;
2: Replacing classifier weight vector W_t^i with c_t^i ;
3: Initializing CVG $G(c_t^i, e(c_t^i, c_m^a))$, according Eq. (5);
4: Reset $ep \leftarrow 0$;
while $ep < ep_{max}$ **do**
1: $ep \leftarrow ep + 1$;
2: Computing L_{inter} according Eq. (9);
3: Updating ϕ, c_t^i;
4: Updating $G(c_t^i, e(c_t^i, c_m^a))$;
end
Output: Backbone ϕ.

pairs. The ablation study in later section proves that the performance of the intra-camera learning stage obtains a great improvement.

3.2. Inter-camera learning

Extracting tracklets' representation. In the inter-camera learning stage, we fuse all of the frames' future as the tracklet's representation c_t^i. The definition of c_t^i is shown in Eq. (3), where $N_{s_t^i}$ is the number of the tracklet frames.

$$c_t^i = \frac{\sum_{n=1}^{N_{s_t^i}} \phi(I_n^{s_t^i})}{N_{s_t^i}}, \quad I_n^{s_t^i} \in s_t^i \tag{3}$$

Building cross-view graph (CVG). We define a local KNN set $\{c_t^i\}_K^m$ of c_t^i, which finds of the nearest K tracklets of c_t^i in camera m. Through merging these local KNN sets, we can get CVG. However, there are lots of noisy links in CVG. In order to reduce them, we apply a **threshold constraint** and a **symmetric constraint** on graph edges. For arbitrary nodes c_t^i and c_m^j on CVG, the former requires the cosine similarity between the nodes c_t^i and c_m^j is larger than the threshold λ, while the latter requires c_t^i and c_m^j must exist in each other's local KNN set. The symmetric constraint can be formulated as:

$$(c_t^i, c_m^j)_K = \{c_m^j \in \{c_t^i\}_K^m \ \& \ c_t^i \in \{c_m^j\}_K^t\} \tag{4}$$

In above equation, if the symmetric constraint can is satisfied, $(c_t^i, c_m^j)_K$ is true; otherwise, $(c_t^i, c_m^j)_K$ is false. If the edge is not satisfied these two constrains, the edge will be removed from CVG. The weight of the edge $e(c_t^i, c_m^j)$ can be summarized as follows :

$$e(c_t^i, c_m^j) = \begin{cases} cos(c_t^i, c_m^j) & if\ cos(c_t^i, c_m^j) > \lambda \ \& \ (c_t^i, c_m^j)_K \\ 1 & if \ \ c_t^i = c_m^j \\ 0 & other \end{cases} \tag{5}$$

where $c_t^i = c_m^j$ denotes the node c_t^i connects itself, *i.e.,* $e(c_t^i, c_t^i)$ is a self-connection. Considering through the SSTT sampling, each person has at most one tracklet in each camera. K is set to 1 in this paper. Through the local KNN set and these two constrains, we can obtain a precise cross-view graph (CVG).

Cross-camera loss. We develop a graph weighted loss as the cross-camera loss to pull the underlying positive pairs close. Firstly, we define a graph neighbor set $N(s_t^i)$ of the tracklet s_t^i:

$$N(s_t^i) = \{(s_m^a, y_m^a) | if\ e(c_t^i, c_m^a) \neq 0\} \tag{6}$$

In fact, $N(s_t^i)$ is a set which contains s_t^i's all local nearest neighbors of all cameras. Though SSTT sampling, we give a pseudo label y_t^i for s_t^i under t-th camera and give a pseudo label y_m^a for s_m^a under m-th camera. 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. Based on this, the pseudo label of s_t^i may be y_m^a under m-th camera, while the pseudo label of s_m^a may be y_t^i under t-th camera. We hope to learn view-invariant representation by pulling these underlying positive pairs close. To this end, we propose the following cross-camera loss:

$$\begin{aligned} l_{ce}(I_n^{s_t^i}, s_m^a) &= -\sum_{j=1}^{M_m} log(\frac{e^{(c_m^j)^T \phi(I_n^{s_t^i})}}{\sum_{k=1}^{M_m} e^{(c_m^k)^T \phi(I_n^{s_t^i})}}) \\ l_{inter}(I_n^{s_t^i}) &= \sum_{N(s_t^i) - s_t^i} l_{ce}(I_n^{s_t^i}, s_m^a), \quad s_m^a \in N(s_t^i) \end{aligned} \tag{7}$$

In above equation, we replace the weight parameters W_t^i of f_t with the corresponding CVG node c_t^i. By doing this, CVG can be fast updated in training process.

Graph weighted cross-camera loss. In Eq. (7), if s_m^a and s_t^i have different IDs, the cross-camera loss will pull negative pairs close. We adopt two strategies to alleviate this problem: 1) using the graph edge's weight as the weighting of the cross-camera loss to reduce the damage of the negative pairs; 2) adding the intra-camera loss as a self-camera constraint item with a constraint weight $\alpha = e(c_t^i, c_t^i)$. The cross-camera loss of the image $I_n^{s_t^i}$ can be redefined as follows:

$$\begin{aligned} l_{inter}(I_n^{s_t^i}) &= \sum_{N(s_t^i) - s_t^i} e(c_t^i, c_m^a) l_{ce}(I_n^{s_t^i}, s_m^a) + \alpha l_{ce}(I_n^{s_t^i}, s_t^i) \\ &= \sum_{N(s_t^i)} e(c_t^i, c_m^a) l_{ce}(I_n^{s_t^i}, s_m^a), \quad where\ \alpha = e(c_t^i, c_t^i) \end{aligned} \tag{8}$$

The multi-branch-classifier loss of the whole mini batch is formulated as follows:

$$l_{inter} = \frac{1}{N_{bs}} \sum_{N_{bs}} l_{inter}(I_n^{s_t^i}), t \in [1, \ldots, T] \quad (9)$$

In Eq. (8), if the model is misled by negative pairs, $l_{ce}(I_n^{s_t^i}, s_t^i)$ will punish it with a large gradient.

CVG's updating. In the above equation, the derivative of c_m^a can be re-emphasized as:

$$\frac{\partial l_{inter}}{\partial c_m^a} = -\sum_{N_{bs}} err(I_n^{s_t^i}) e(c_t^i, c_m^a) \phi(I_n^{s_t^i}) \quad (10)$$

with:

$$err(I_n^{s_t^i}) = \mathbf{1}(y_m^a == j) - \frac{e^{(c_m^j)^T \phi(I_n^{s_t^i})}}{\sum_{k=1}^{M_m} e^{(c_m^k)^T \phi(I_n^{s_t^i})}} \quad (11)$$

The updating of CVG's node c_m^a can be formulated as:

$$c_m^a \leftarrow c_m^a - \eta \sum_{N_{bs}} err(I_n^{s_t^i}) e(c_m^a, c_t^i) \phi(I_n^{s_t^i}), t \in [1, \ldots, T] \quad (12)$$

In above equation, η denotes the learning rate and M_m is the total number of tracklets in camera m. According to Eq. (12), the updating of c_t^i makes full use of underlying positive pairs from all camera views. This measure pulls underlying positive pairs close and encourages CVG finding more cross-view underlying positive pairs.

4. Experiment

4.1. Experimental Setup

Datasets and evaluation protocol. All experiments are evaluated on four image RE-ID datasets (Market-1501 [46], DukeMTMC-ReID [27, 47], CUHK03-detected [20], MTMS17 [38]) and three video RE-ID datasets (Mars [45], Prid2011 [11], iLIDS-Video [37]). The ablation studies are mainly conducted on Market-1501 [46] and Mars [45] 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 labelled images and detected images separately, which needs to repeat 20 times of random 1,367/100 training/testing identity splitting and report the averaged results.

Pseudo label assignment. We follow the experiments settings and tracklet sampling methods of TAUDL [17] and UTAL [18]. 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 sampling 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 are belong to a single tracklet. Then, we randomly assign a unique pseudo label to each tracklet for each camera.

Table 1: Dataset statistics and training/testing splitting

Dataset	ID	Cam	Track	Tain	Test	Images
iLDS-VID	300	2	600	150	150	43,800
PRID2011	178	2	354	89	89	38,466
MARS	1,261	6	20,478	625	636	1,191,003
Market	1,501	6	0	751	750	32,668
Duke	1,812	8	0	702	1,110	36,411
MSMT17	4,101	15	0	1,041	3,060	126,441
CUHK03	1,467	2	0	1,367	100	14,096

"Market", "Duke" and "CUHK03" denote Market-1501, DukeMTMC-ReID and CUHK03-detected datasets respectively.

4.2. Implement details

The structure of the backbone is shown in Figure 2.The training images are resized to 256×128. In order to balance the model learning speed over different cameras, we adopt an equably sampling strategy, *i.e.,* randomly sampling 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 $3.5e^{-4}$ and decaying 0.1 at the 40-th and 60-th epoch. The hyper-parameter λ is set to 0.65. The total training epoch is 80 for both the intra-camera and inter-camera learning stage.

4.3. Ablation Study

BN analysis. As shown in Figure 3, after adding a BN layer for both the supervised algorithm and unsupervised algorithm, the average similarity of negative samples becomes 17 times smaller than that of positive samples in

Table 2: The ablation studies of the BN.

strategies	Market-1501		Mars	
metric(%)	mAP	Rank 1	mAP	Rank 1
R	27.9	47.3	26.0	41.1
R+BN	55.6	78.9	33.2	53.4
R+embed	54.8	77.5	35.1	55.1
R*	73.5	88.2	47.6	62.2
R+BN*	77.1	91.5	51.7	67.6
R+embed*	77.9	91.2	55.7	69.6

[1] $*$ denotes the supervised algorithm.
[2] "R" denotes only use the Resnet-50 backbone;
[3] "R+BN" denotes adding a BN after the backbone;
[4] "embed" denotes adding an embedding block after the backbone.

Table 3: The performance of different λ

	Market		DukeMTMC		CUHK03		MSMT17		Mars		Prid2011		iLIDS-VID	
metric(%)	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	R1	R5	R1	R5
0.55	67.5	85.5	54.2	74.8	70.5	59.6	20.2	46.0	38.7	58.1	71.9	92.1	54.0	74.0
0.6	68.9	86.3	55.2	74.3	69.4	57.2	21.7	50.2	40.5	59.9	79.8	93.3	51.3	72.7
0.65	70.3	87.2	53.3	75.0	68.2	56.5	21.7	49.5	39.3	58.1	80.9	94.4	57.3	72.0
0.7	71.0	87.9	55.7	75.7	63.4	51.0	20.9	47.3	37.8	57.7	77.5	92.1	47.3	70.0
0.75	69.3	86.3	55.1	75.0	61.6	48.4	21.3	49.2	35.5	54.5	70.8	91.0	48.0	69.3

Table 4: The performance of the intra-learning stage and inter-learning stage

	Market		DukeMTMC		CUHK03		MSMT17		MARS		Prid2011		iLIDS-VID	
metric(%)	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	R1	R5	R1	R5
intra	54.8	77.5	52.5	72.6	56.3	42.2	19.7	45.7	35.1	55.1	65.2	86.5	42.7	74.0
inter($\lambda = 0.65$)	70.3	87.2	53.3	75.0	68.2	56.5	21.7	49.5	39.3	58.1	80.9	94.4	57.3	72.0
improvement	+15.5	+9.7	+0.8	+2.4	+11.9	+14.3	+2.0	+3.8	+4.2	+3.0	+15.7	+7.9	+14.6	-2.0

the unsupervised algorithm, while it becomes the 355 times smaller in the supervised algorithm. It indicates the BN layer helps to make the positive and negative samples easier to be discriminated. Particularly, the unsupervised tracklet based algorithm obtains a huge promotion. As in previous work ([12, 29]), BN may help the deep network converge faster and we find that the faster convergence helps better distinguish negative pairs. We compare two BN structures in Table 2: 1) adding a batch normalization layer at the top of the Resnet-50 backbone (R+BN); 2) applying an embedding structure (BN-Dropout-FC-BN) after the Resnet-50 backbone. The performances of the two structure are shown in Table 2. R+BN achieves a better result on Market-1501, while the embedding block performs better on MARS. To avoid overfitting, we adopt the embedding block in this paper, since the embedding block includes a dropout layer.

Table 5: Effect of the self-camera constraint item.

	Market-1501		MARS	
Metric(%)	mAP	R 1	mAP	R 1
base[1]	54.8	77.5	53.1	55.1
w/o self-cam[2]	65.8	83.7	36.7	53.4
w self-cam[2]	71.0	87.9	40.5	59.9

[1] Baseline model of intra-camera learning stage.
[2] Self-camera Constraint item.

Table 6: Robust analysis of noisy tracklets on MARS

Rate	ID duplication			Mislabeling		
Metric(%)	mAP	R1	R5	mAP	R1	R5
0%	35.1	55.1	74.6	35.1	55.1	74.6
20%	31.9	54.1	68.9	33.4	53.3	67.2
50%	30.1	51.4	66.3	26.8	45.7	61.4
100%	27.2	48.8	64.2	-	-	-

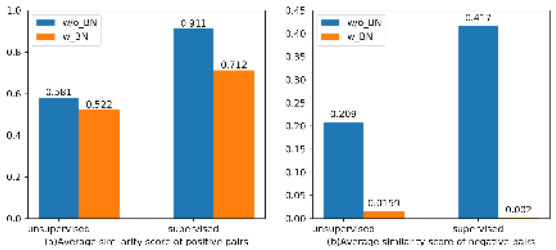

Figure 3: (a) and (b) show the average similarity score of positive pairs and negative pairs on Market-1501, respectively. The average similarity score of negative pairs declines obviously after using BN for both of the supervised and unsupervised training.

Threshold λ analysis. The threshold λ is important for the initialization and updating of CVG. The precision scores and recall scores with different λ are shown in Figure 4 . With λ increasing from 0.1 to 0.9, the precision is closing to 1 while the recall score is declining to 0. As shown in Figure 4, the precision score is greatly improved after using the symmetric constraint. It proves that our symmetric constraint strategy is effective. There is a good trade-off between recall and precision, when λ is set between 0.55 and 0.75. Therefore, in the Table 3, we evaluate the performance of λ from $\{0.55, 0.6, 0.65, 0.7, 0.75\}$. From the average Rank-1, $\lambda = 0.65$ achieves the best performance, and $\lambda = 0.75$ performs worst.

Self-camera constraint item analysis. We introduce a self-camera constraint item into the cross-camera loss (Eq. (8)), to alleviate the misleading by noisy associations. The ablation study of the self-camera constraint item is shown in Table 5. The self-camera constraint item improves 5.2% mAP and 4.2% rank-1 in Market-1501, while improving 3.8% mAP and 5.4% rank-1 in MARS. Particularly

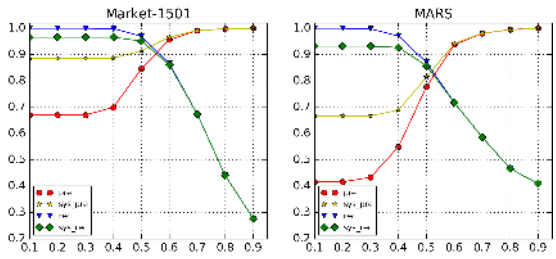

Figure 4: Precision scores and recall scores of different λ. In this figure,"pre" denotes the precision. "rec" denotes the recall. "sys_pre" and "sys_rec" respectively denote the precision and recall of using the symmetry condition. The horizontal axis is the value of the λ.

in the MARS, without the self-camera constraint item, the inter-camera learning even brings down the performance of the intra-camera learning stage.

Effectiveness of the cross-camera loss. The inter-camera training stage encourages the model to learn the view-invariant representations. The performance of the intra-camera learning and inter-camera learning are shown in Table 4. We can observe that the inter-camera stage averagely improves 7.55% rank-1 for image RE-ID datasets, and averagely improves 11.1% rank-1 for video RE-ID datasets. This demonstrates the effective of CVG and the cross-camera loss.

Robust analysis of intra-camera stage. 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 mislabelling often occur in practice. The ID duplication is that the tracklets of the same person are given different pseudo labels. While the mislabelling is assigned the tracklets of different persons with the same pseudo labels. The base of UGA is the intra-camera learning stage. To evaluate the robust of this stage, we simulate noisy tracklets in these two situations. **For the ID duplication situation**, we randomly select a part (20%, 50%, 100%) of persons per camera to create the ID duplication, while the remaining IDs still sample one tracklet. In Table 6, we can see that 20% of the persons have ID duplication, the model of the intra-camera stage declines by 1% on rank-1; when 50% of the persons have ID duplication, the model of the intra-camera stage declines by 3.7% on rank-1; when all the persons have ID duplication, the model of the intra-camera stage still achieves 48.8% rank-1 and 27.2% mAP. The model of the intra-camera stage is not very sensitive to the ID duplication noise. **For the mislabelling situation**, we randomly merge a portion (20%, 50%) of tracklets to simulate the mislabelling situation. When merging 20% of all the tracklets, rank-1 is decreased by 1.8%; when merging with 50% of all the tracklets, rank-1 is decreased by 9.4%. Under the influence of two kinds of noise, the intra-camera learning stage model still achieves a competitive performance.

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

We compare our UGA with some state-of-the-art unsupervised person RE-ID methods, specifically comparing with four similar unsupervised graph based methods. The performances of these methods are shown in Table 7 and Table 8.

Image person RE-ID datasets. Table 7 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). The proposed UGA outperforms all these approaches. Specifically, UGA averagely outperforms the second by 9.6% on Rank-1 accuracy and 16.8% on mAP, respectively. Both of the adaptation methods and cluster methods rely on source domain adaptation, specifically cluster method is inefficient on the large dataset (*i.e.*, MSMT17) since it will spend much time on offline data clustering. Comparing with them, UGA has better generalization ability, since UGA does not need source domain pre-training and the association progress (CVG) can be updated online.

Video person RE-ID datasets. We compare the proposed UGA on three video person RE-ID datasets with several state-of-the-art approaches in Table 8. The proposed UGA outperforms all the state-of-the-art methods on iLIDS-VID, 15.6% higher than the second (SMP) on Rank-1. On Prid2011, UGA is also competitive and even reaches 100% on Rank-20. On MARS, our approach does not perform as good as EUG and BUC. However, as shown in BUC [23] and EUG [39], BUC is sensitive to the hyper-parameter and merging times, while EUG is sensitive to the enlarging factors. When the enlarging factors changing, the rank-1 of EUG declines from 62.67% to 42.77%. The Rank-1 of UGA varies from 59.9% to 54.5% with the λ changing. Comparatively, UGA is more robust to the hyper-parameter.

Comparison with the unsupervised graph based methods. We compare our UGA with the existed graph based work (*i.e.*, TUADL [17], UTAL [18], RACE [40] and ECN [50]) in Table 7 and Table 8. UGA averagely outperforms TAUDL by (17.4% on Rank-1, 21.3% on mAP) in image person RE-ID datasets and (25.4% on Rank-1, 16.6% on Rank-5) in video person RE-ID datasets. UGA outperforms UTAL by (12.3% on Rank-1, 16.8% on mAP) in image person RE-ID datasets and (18.9% on Rank-1, 10.4% on Rank-5) in video person RE-ID datasets. In addition, both of TAUDL and UTAL matches the positive pairs in the

Table 7: Comparing UGA with the state-of-the-art methods on the image person RE-ID dataset

Dataset	Reference	Method	Market1501		DukeMTMC-ReID		CUHK03		MSMT17	
metric			mAP	Rank 1	mAP	Rank 1	mAP	Rank 1	mAP	Rank 1
HHL [49]	ECCV'18	GAN	31.4	62.2	27.2	46.9	-	-	-	-
SPGAN [4]	CVPR'18	GAN	22.8	51.5	22.3	41.1				
SPGAN+LMP [4]	CVPR'18	GAN	26.7	57.7	26.2	46.4	-	-	-	-
TJ-AIDL [36]	CVPR'17	adaptation	26.5	58.2	23.0	44.3	-	-	-	-
BUC [23]	AAA'19	cluster	38.3	66.2	27.5	47.4	-	-	-	-
CAMEL [43]	ICCV'17	cluster	26.3	54.5	-	-	-	39.4	-	-
PUL [5]	ToMM'18	cluster	20.1	44.7	16.4	30.4	-	-	-	-
CDS [16]	ICME'19	cluster	39.9	71.6	42.7	67.2	-	-	-	-
TAUDL [17]	ECCV'18	tracklet	41.2	63.7	43.5	61.7	31.2	44.7	12.5	28.4
UTAL [18]	TPAMI'19	tracklet	46.2	69.2	44.6	62.3	42.3	56.3	13.1	31.4
ECN [50]	CVPR'19	adaptation	43.0	75.1	40.4	63.3	-	-	10.2	30.2
UGA(ours)	This work	tracklet	70.3	87.2	53.3	75.0	68.2	56.5	21.7	49.5

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

Table 8: Comparing UGA with the state-of-the-art methods on the video person RE-ID dataset.

Datasets	Reference	PRID2011			iLIDS-VID			MARS			
Metric(%)		R1	R5	R20	R1	R5	R20	R1	R5	R20	mAP
SMP [24]	ICCV'17	80.9	95.6	99.4	41.7	66.3	80.7	23.9	35.8	44.9	10.5
DGM+MLAPG [41]	ICCV'17	73.5	92.6	99.0	37.1	61.3	82.0	24.6	42.6	57.2	11.8
DGM+IDE [41]	ICCV'17	56.4	81.3	96.4	36.2	62.8	82.7	36.8	54.0	68.5	21.3
DASy [1]	ECCV'18	43.0	-	-	56.5	-	-	-	-	-	-
GRDL [13]	ECCV'16	41.6	76.4	89.9	25.7	49.9	77.6	19.3	33.2	46.5	9.56
DTW [26]	PR'17	41.7	67.1	90.1	31.5	62.1	82.4	-	-	-	-
BUC [23]	AAAI'19	-	-	-	-	-	-	61.1	75.1	80.0	38.0
EUG(p=0.05) [39]	CVPR'18	-	-	-	-	-	-	62.7	74.9	82.6	42.5
RACE [40]	ECCV'18	50.6	79.4	91.8	19.3	39.3	68.7	43.2	57.1	67.6	24.5
TAUDL [17]	ECCV'18	49.4	78.7	98.9	26.7	51.3	82.0	43.8	59.9	72.8	29.1
UTAL [18]	TPAMI'19	54.7	83.1	96.2	35.1	59.0	83.8	49.9	66.4	77.8	35.2
UGA(ours)	This work	80.9	94.4	100	57.3	72.0	87.3	58.1	73.4	81.4	39.3

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

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 UGA can be implemented on one 1080-Ti, since CVG can be stored in CPU memory. Different from RACE [40] merging the underlying positive tracklets directly, UGA uses the cross-camera loss and CVG to associate tracklets. It is more robust to noisy associations. Due to this, UGA easily achieves the higher performance than RACE. Comparing with ECN, UGA averagely outperforms ECN by (14.4% on Rank-1, 16.1% on mAP) in image person RE-ID datasets. Because ECN is simply apply a KNN graph to associate the underlying positive samples, while UGA uses the more precise graph (CVG) to associate the underlying positive pairs.

5. Conclusion

In this paper, we have proposed a novel yet effective Unsupervised Graph Association (UGA) approach to address the unsupervised person RE-ID problem. The core ideas of UGA are finding more underlying correct associations and avoiding the damage of noisy associations. To that end, we mainly adopt an embedding block, a cross-view graph (CVG) mining strategy and a graph weighted cross-camera loss. Experiments on four image RE-ID dataset and three video RE-ID dataset demonstrate the superiority of UGA.

Acknowledgements

This work was supported by the Chinese National Natural Science Foundation Projects #61672521, #61806203, #61876178, #61872367 and #61572501.

References

[1] Slawomir Bak, Peter Carr, and Jean-Francois Lalonde. Domain adaptation through synthesis for unsupervised person re-identification. *arXiv preprint arXiv:1804.10094*, 2018.

[2] Xiaobin Chang, Timothy M Hospedales, and Tao Xiang. Multi-level factorisation net for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2109–2118, 2018.

[3] Dapeng Chen, Dan Xu, Hongsheng Li, Nicu Sebe, and Xiaogang Wang. Group consistent similarity learning via deep crf for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8649–8658, 2018.

[4] Ye Qixiang Kang Guoliang Yang Yi Jiao Jianbin Deng Weijian, Zheng Liang. 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*, pages 994–1003, 2018.

[5] Hehe Fan, Liang Zheng, Chenggang Yan, and Yi Yang. Unsupervised person re-identification: Clustering and fine-tuning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(4):83, 2018.

[6] Junyu Gao, Tianzhu Zhang, and Changsheng Xu. Graph convolutional tracking. In *CVPR*, 2019.

[7] Junyu Gao, Tianzhu Zhang, and Changsheng Xu. I know the relationships: Zero-shot action recognition via two-stream graph convolutional networks and knowledge graphs. In *AAAI*, 2019.

[8] Chen Gong, Tongliang Liu, Dacheng Tao, Keren Fu, Enmei Tu, and Jie Yang. Deformed graph laplacian for semisupervised learning. *IEEE transactions on neural networks and learning systems*, 26(10):2261–2274, 2015.

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[11] Martin Hirzer, Csaba Beleznai, Peter M Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *Scandinavian conference on Image analysis*, pages 91–102. Springer, 2011.

[12] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[13] Elyor Kodirov, Tao Xiang, Zhenyong Fu, and Shaogang Gong. Person re-identification by unsupervised l_1 graph learning. In *European conference on computer vision*, pages 178–195. Springer, 2016.

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[15] Laura Lealtaixe, Anton Milan, Ian D Reid, Stefan Roth, and Konrad Schindler. Motchallenge 2015: Towards a benchmark for multi-target tracking. *arXiv: Computer Vision and Pattern Recognition*, 2015.

[16] Jinlin Wu. Shengcai Liao. Zhen Lei. Xiaobo Wang. Yang Yang. Stan Z. Li. Clustering and dynamic sampling for unsupervised domain adaptation in person re-identification. In *IEEE International Conference on Multimedia and Expo (ICME)*, 2018.

[17] Minxian Li, Xiatian Zhu, and Shaogang Gong. Unsupervised person re-identification by deep learning tracklet association. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 737–753, 2018.

[18] Minxian Li, Xiatian Zhu, and Shaogang Gong. Unsupervised tracklet person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2019.

[19] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 152–159, 2014.

[20] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014.

[21] Yu-Jhe Li, Fu-En Yang, Yen-Cheng Liu, Yu-Ying Yeh, Xiaofei Du, and Yu-Chiang Frank Wang. 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*, pages 172–178, 2018.

[22] Wenqi Liang, Guangcong Wang, Jianhuang Lai, and Junyong Zhu. M2m-gan: Many-to-many generative adversarial transfer learning for person re-identification. *arXiv preprint arXiv:1811.03768*, 2018.

[23] Yutian Lin, Xuanyi Dong, Liang Zheng, Yan Yan, and Yi Yang. A bottom-up clustering approach to unsupervised person re-identification. In *AAAI Conference on Artificial Intelligence*, volume 2, 2019.

[24] Zimo Liu, Dong Wang, and Huchuan Lu. Stepwise metric promotion for unsupervised video person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2429–2438, 2017.

[25] Yucen Luo, Jun Zhu, Mengxi Li, Yong Ren, and Bo Zhang. Smooth neighbors on teacher graphs for semi-supervised learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8896–8905, 2018.

[26] Xiaolong Ma, Xiatian Zhu, Shaogang Gong, Xudong Xie, Jianming Hu, Kin Man Lam, and Yisheng Zhong. Person re-identification by unsupervised video matching. *Pattern Recognition*, 65(C):197–210, 2016.

[27] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision*, pages 17–35. Springer, 2016.

[28] Ergys Ristani, Francesco Solera, Roger S Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data

set for multi-target, multi-camera tracking. *european conference on computer vision*, pages 17–35, 2016.

[29] Yantao Shen, Hongsheng Li, Shuai Yi, Dapeng Chen, and Xiaogang Wang. Person re-identification with deep similarity-guided graph neural network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 486–504, 2018.

[30] Hailin Shi, Yang Yang, Xiangyu Zhu, Shengcai Liao, Zhen Lei, Weishi Zheng, and Stan Z Li. Embedding deep metric for person re-identification: A study against large variations. In *European conference on computer vision*, pages 732–748. Springer, 2016.

[31] Yifan Sun, Liang Zheng, Weijian Deng, and Shengjin Wang. Svdnet for pedestrian retrieval. *ICCV*, pages 3820–3828, 2017.

[32] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 480–496, 2018.

[33] Zichang Tan, Yang Yang, Jun Wan, Hanyuan Hang, Guodong Guo, and Stan Z. Li. Attention based pedestrian attribute analysis. *IEEE transactions on image processing*, 2019.

[34] Guangcong Wang, Jianhuang Lai, Peigen Huang, and Xiaohua Xie. Spatial-temporal person re-identification. *national conference on artificial intelligence*, 2019.

[35] Guan'an Wang, Yang Yang, Jian Cheng, Jinqiao Wang, and Zengguang Hou. Color-sensitive person re-identification. In *IJCAI 2019: 28th International Joint Conference on Artificial Intelligence*, 2019.

[36] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2275–2284, 2018.

[37] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by video ranking. In *European Conference on Computer Vision*, pages 688–703. Springer, 2014.

[38] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 79–88, 2018.

[39] Yu Wu, Yutian Lin, Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. 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*, pages 5177–5186, 2018.

[40] Mang Ye, Xiangyuan Lan, and Pong C Yuen. Robust anchor embedding for unsupervised video person re-identification in the wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 170–186, 2018.

[41] Mang Ye, Andy J Ma, Liang Zheng, Jiawei Li, and Pong C Yuen. Dynamic label graph matching for unsupervised video re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5142–5150, 2017.

[42] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Deep metric learning for person re-identification. In *2014 22nd International Conference on Pattern Recognition*, pages 34–39. IEEE, 2014.

[43] Hong-Xing Yu, Ancong Wu, and Wei-Shi Zheng. Cross-view asymmetric metric learning for unsupervised person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 994–1002, 2017.

[44] Zhipeng Zhang and Houwen Peng. Deeper and wider siamese networks for real-time visual tracking. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[45] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. Mars: A video benchmark for large-scale person re-identification. In *European Conference on Computer Vision*, pages 868–884. Springer, 2016.

[46] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015.

[47] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. *international conference on computer vision*, pages 3774–3782, 2017.

[48] Zhedong Zheng, Liang Zheng, and Yi Yang. A discriminatively learned cnn embedding for person reidentification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(1):13, 2018.

[49] Zhun Zhong, Liang Zheng, Shaozi Li, and Yi Yang. Generalizing a person retrieval model hetero-and homogeneously. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–188, 2018.

[50] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 598–607, 2019.

[51] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. Camstyle: A novel data augmentation method for person re-identification. *IEEE Transactions on Image Processing*, 28(3):1176–1190, 2018.

[52] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. Camstyle: A novel data augmentation method for person re-identification. *IEEE Transactions on Image Processing*, 28(3):1176–1190, 2019.