Pattern Recognition 93 (2019) 430–441

Contents lists available at ScienceDirect

Pattern Recognition

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

Deep asymmetric video-based person re-identification

Jingke Meng[a,d], Ancong Wu[c], Wei-Shi Zheng[a,b,d,*]

[a] School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China
[b] Peng Cheng Laboratory, Shenzhen, China
[c] School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou, China
[d] Key Laboratory of Machine Intelligence and Advanced Computing Sun Yat-sen University, Ministry of Education, China

ARTICLE INFO

Article history:
Received 14 August 2018
Revised 4 April 2019
Accepted 9 April 2019
Available online 4 May 2019

Keywords:
Person re-identification
Visual surveillance

ABSTRACT

In this paper, we investigate the problem of video-based person re-identification (re-id) which matches people's video clips across non-overlapping camera views at different time. A key challenge of video-based person re-id is a person's appearance and motion would always display differently and take effects unequally at disjoint camera views due to the change of lighting, viewpoint, background and etc., which we call the "view-bias" problem. However, many previous video-based person re-id approaches have not quantified the importance of different types of features at different camera views, so that the two types of important features (i.e. appearance and motion features) do not collaborate effectively and thus the "view-bias" problem remains unsolved. To address this problem, we propose a ***Deep Asymmetric Metric learning* (DAM)** method that embeds a proposed *asymmetric distance metric learning loss* into a two-stream deep neural network for jointly learning view-specific and feature-specific transformations to overcome the "view-bias" problem in video-based person re-id. As learning these view-specific transformations become expensive when there are large amount of camera views, a clustering-based DAM method is developed to make our DAM scalable. Extensive evaluations have been carried out on three public datasets: PRID2011, iLIDS-VID and MARS. Our results verify that learning view-specific and feature-specific transformations are beneficial, and the presented DAM has empirically performed more effectively overall for video-based person re-id on challenging benchmarks.

1. Introduction

Person re-identification (re-id) matches persons across non-overlapping camera views at different time. Given a probe image, we match it against a set of gallery images, which may suffer from the changes of illumination, camera viewpoint, background and occlusions. This leads to serious visual ambiguity and appearance variation and makes image-based person re-id a challenging problem. Many methods have been developed from either extracting hand-craft invariant features [1–6] or learning discriminative matching models [4,7–32] to end-to-end deep learning methods [33–43]. However, all these image-based models assume that one image frame or several are selected in advanced.

More recently, there have been attempts of processing people's video clips directly. On the one hand, it is expected to select or quantify features extracted from video frames automatically; on the other hand, the extra dynamic information is exploited from a short clip of a person for helping to alleviate the influence of occlusion and background in person re-id. In general, describing a person's video can naturally be attributed to spatial and temporal cues. The spatial part carries scenes and appearance information (e.g. clothing color, height, and shape) of a person, and the temporal part in the form of motion across frames conveys a person's movement, which is complementary to the spatial part. So far, several video-based methods have been proposed [44–57], and most of them use the temporal information as an auxiliary supplement to the spatial one and finally get a spatial-temporal representation.

A key challenge in video-based person re-id is how to quantify the appearance and motion information. Always, a person's image sequences are presented differently at disjoint camera views because the environment and camera deployment can be entirely different (as shown in Fig. 1), and thus the appearance (or motion) feature can be quite different across camera views as well. Besides, as illustrated in Fig. 2, appearance feature and motion feature describe different aspects of a person, and we can also find that the motion feature can be much more useful when different people look similar (e.g. wearing similar color clothes like person

* Corresponding author at: School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China.
E-mail addresses: mengjke@mail2.sysu.edu.cn (J. Meng), wuancong@mail2.sysu.edu.cn (A. Wu), wszheng@ieee.org (W.-S. Zheng).

https://doi.org/10.1016/j.patcog.2019.04.008
0031-3203/

Fig. 1. The illustration of cross-view video pair instances. The same person from different camera views look quite different because the environment and camera deployment can be entirely different.

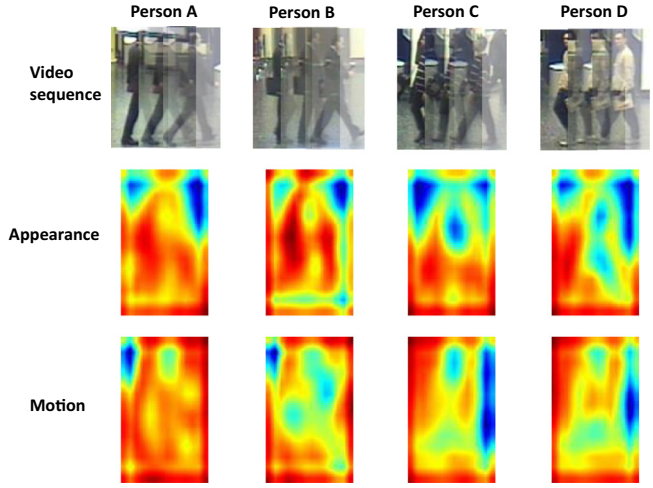


Fig. 2. The illustration of video feature instance. The first row is the video frame sequence, the bottom two rows are the corresponding appearance feature map and motion feature map which are extracted from the last convolution layer. The person A and person B are similar in appearance, but their motion features are clearly different. The person C and person D have similar movement patterns but their appearance features are different.

A and person B in Fig. 2), while the appearance feature seems more effective if different people have similar movement patterns (e.g. person C and person D in Fig. 2).

Hence, it is found that (1) the same type of feature could perform differently at different camera views (i.e. view-specific bias), and (2) the contributions of different types of features are not equal as well (i.e. feature-specific bias). In this work, we refer the above problems as the "view-bias" problem. Although many previous video-based person re-id approaches [44–56,58] have taken both appearance and motion information into consideration, they have not considered the view-specific transformations of different types of features at different camera views, i.e. they are not able to tell whether a feature should contribute differently in different camera views and how all features collaborate. To this end, the two crucial types of features do not collaborate effectively and thus the "view-bias" problem remains unsolved.

Essentially, a key problem of previous approaches for video-based person re-id [47–49,51–54] are to learn universal transformation at all camera views for the same type of feature. In particular, given a pair of sample representations $\mathbf{x}_i$ and $\mathbf{x}_j$ from different camera views, the commonly employed Mahalanobis distance metric in video-based person re-id models learns universal feature transformation for data samples at all camera views form the symmetric distance metric below:

$$d(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^T M (\mathbf{x}_i - \mathbf{x}_j) = \|U^T \mathbf{x}_i - U^T \mathbf{x}_j\|_2^2 \quad (1)$$

where M is the positive semidefinite matrix and can be factorized into $M = UU^T$, and U is the feature transformation matrix which projects $\mathbf{x}_i$ and $\mathbf{x}_j$ into a shared feature subspace and the Euclidean distance is calculated afterwards. The universal transformation U is not enough for modeling view-bias problem well.

In this work, we overcome the above problem by developing an asymmetric-based learning method. Specifically, we design to learn the view-specific transformation for the same type of feature at different camera views based on the following asymmetric distance metric:

$$d(\mathbf{x}_i^v, \mathbf{x}_j^c) = \|U_v^T \mathbf{x}_i^v - U_c^T \mathbf{x}_j^c\|_2^2, \quad (2)$$

where $\mathbf{x}_i^v$ denotes the ith sample from the vth view, and U_v is the corresponding view-specific transformation matrix. In this way, we can learn the *view-specific transformation* to alleviate the view-bias problem to a certain extent.

Moreover, since appearance feature and motion feature take effects unequally at different camera views, we learn the feature transformation for appearance and motion features separately, based on the following modified asymmetric distance metric

$$d(\mathbf{x}_i^v, \mathbf{x}_j^c) = \alpha \|U_{v,a}^T \mathbf{x}_{i,a}^v - U_{c,a}^T \mathbf{x}_{j,a}^c\|_2^2 + (1-\alpha)\|U_{v,m}^T \mathbf{x}_{i,m}^v - U_{c,m}^T \mathbf{x}_{j,m}^c\|_2^2, \quad (3)$$

where the extra notations "a" and "m" are used to indicate the type of transformation, i.e. $U_{v,a}$ is the transformation on appearance feature of the vth view, $U_{v,m}$ is the transformation on motion feature of the v-th view. As such, we also call $U_{v,a}$ and $U_{v,m}$ not only the view-specific transformation but also the *feature-specific transformation*.

To this end, we develop a novel ***Deep Asymmetric Metric learning* (DAM)** Method for video-based person re-id which embeds the above asymmetric distance metric into a two-stream neural network so as to perform an end-to-end joint learning of view-specific transformation and feature-specific transformation. By such an asymmetric learning, we are also able to learn the transformations of different features at different camera views so that they are selectively used and quantified at different views. During learning view-specific and feature-specific transformations, the features of persons from different cameras are projected into a latent subspace of the same dimension using these projection matrices. We quantify these projections by concerning the relative comparison between the maximum distance of positive pairs (intra-class person samples) and the distance of related negative pairs (inter-class person samples), rather than comparing the negative pair with each of the relevant positive pairs, so as to reduce the complexity on quantifying large amount of relative comparisons between them.

Also, we notice that it could not be tractable to learn the view-specific metric when there are thousands of camera views. To alleviate this situation, we develop a clustering-based DAM method. The developed clustering-based DAM attempts to cluster these camera views according to their samples using the agglomerative hierarchical clustering algorithm to get the clusters firstly, and then we learn the cluster-specific transformation through our DAM afterwards.

Extensive evaluations are carried out on three public datasets: PRID 2011, iLIDS-VID and MARS. The results demonstrate that our approach outperforms or sometimes works comparably with a

wide range of state-of-the-art video-based person re-id methods. And, we empirically find that learning feature-specific transformation rather than straightforward embedding the temporal motion information in the learning process of spatial appearance makes benefit, and also the usefulness of each part and how to balance them are evaluated on video-based person re-id.

2. Related works

2.1. Video-based person re-id

Always, based on the tracklets extracted by the tracking algorithms [59–61], we need to match pedestrians across disjoint camera views in the video-based person re-id. Recently, an important approach in video-based person re-id is to take the motion information from a short video clip of a person into consideration for matching people across disjoint camera views [44–56,58]. The early stage of this approach has seen a lot of development under a two-stage framework [44–46] which firstly extracts handcraft spatial-temporal representations (e.g. HOG3D [62], STFV3D [63], TAPR [55]) as video sequence descriptor and then learns a discriminative distance metric.

More recently, the deep approach that unifies feature extraction and distance metric learning is also explored for video-based person re-id. Among them, the recurrent neural network is widely used to capture the dynamic temporal information of a video [48,49,51–53,64]. Besides, the attention mechanism is adopted in [52,53,65–68] to automatically pick out the most discriminative information hidden in the video sequence. Furthermore, while two-stream convolutional networks have been widely used for action recognition [69], it is also applied on the video-based person re-id [47,56,70] to learn and fuse the RGB and optical flow information.

Albeit successfully applying the above algorithms as reported in literature, existing video-based person re-id methods are all symmetric which take all the camera views equally and essentially learn the same transformation for each type of feature at different camera views and do not explicitly identify the different contributions of appearance feature and motion feature, making the "view-bias" problem remained unsolved. In contrast, our proposed model is a deep asymmetric distance metric learning so as to learn view-specific transformation for the same type of feature at different camera views. Furthermore, as compared to [47–49,51–54], we learn feature-specific transformation for each type of feature rather than learning transformation on the concatenated features of different types, and this enables our algorithm to directly evaluate the different effectiveness of appearance feature and motion feature for identifying a person. Moreover, we design the asymmetric loss in order to concern the relative distance comparison between the maximum distance of positive pairs and the distance of related negative pairs, rather than comparing the negative pair with each of the relevant positive pairs, and this would notably reduce the computational complexity of our model.

2.2. Deep distance metric learning

Most conventional models are known for learning a subspace or discriminative metric which mostly maximizes cross-view inter-person distance and minimizes cross-view intra-person distance [4,7–13,15,16,18–27,41–43], and among them the relative distance comparison or the triplet loss based methods are widely explored for person re-id [7–9,13,18,20–22,25,26]. There are also some multi-view learning methods [71–73] adopted to solve the cross-view person re-id problem, such as the CCA-based learning algorithms [74,75]. While previous distance metric learning models are mostly second-order based, it has seen a lot of development on the deep approach [35,36,76–85]. The deep neural networks are able to extend distance metric learning from lower-order distance metric function to higher-order ones so as to learn highly nonlinear data variations and sometimes jointly learn features from raw data as well. Our work is more related to these recent deep distance metric learning methods. These previous deep distance metric learning works are symmetric when they are applied for person re-id, namely learning universal transformation function for data samples from all camera views. In comparison, our proposed deep asymmetric metric learning (DAM) is the first deep distance metric based on the asymmetric learning idea for learning view-specific and feature-specific transformation functions jointly for different views and different feature types. In such a way, our DAM is able to identify the importance of different feature types and make them collaborate more efficiently to overcome the view-bias problem, and this is not realized in the previous deep distance metric learning for video-based person re-id.

Our asymmetric learning model is related to several recent asymmetric learning methods for image-based distance learning models for person re-id, such as the work proposed in [20] and [21]. The differences are three folds: (1) the proposed DAM is a deep asymmetric metric learning method in an end-to-end way, while it is not in [20] and [21]; (2) the proposed DAM realizes a joint learning of view-specific and feature-specific metric rather than learning the asymmetric distance metrics for different feature types (i.e. the visual appearance feature and the motion feature) individually; (3) the previous asymmetric modeling cannot be scalable to a large number of camera views, and thus a clustering-based DAM model is introduced to alleviate this problem.

3. Deep asymmetric metric learning method

Overall, we form a deep asymmetric metric learning model for jointly learning view-specific and feature-specific transformations under the asymmetric metric formula (Eq. (3)). The proposed asymmetric distance metric will be embedded into the widely used two-stream neural network, and the demonstration of our deep asymmetric metric learning process on two camera views network is shown in Fig. 3. Let the input of be denoted by $\mathbf{p}_i^v = \{\mathbf{I}_{i,a}^v, \mathbf{S}_{i,m}^v\} = \{\{\mathbf{I}_{i,a,t}^v\}_{t=1}^{N_i}, \{\mathbf{S}_{i,m,t}^v\}_{t=1}^{N_i}\}$, where the $\mathbf{I}_{i,a,t}^v$ represents the tth single RGB image which describes the appearance information of the ith video sample under the vth ($v = 1, 2, \ldots, V$) camera view and the associated $\mathbf{S}_{i,m,t}^v$ is the optical flow sequence at time t which conveys the movement or gait information.

By an end-to-end system, the feature representation and transformation matrices can be learned at the same time. We assume that the to-learn feature extractor functions f^a and f^m are spatial CNN sub-network and temporal CNN sub-network respectively. We denote the appearance feature of single frame at time t as $\mathbf{x}_{i,a,t}^v = f^a(\mathbf{I}_{i,a,t}^v)$ and the motion feature of optical flow sequence at time t as $\mathbf{x}_{i,m,t}^v = f^m(\mathbf{S}_{i,m,t}^v)$. To obtain stable feature representation, we express the feature of a video sample by average pooling of features of all frames from that video, that is the corresponding video appearance feature $\mathbf{x}_{i,a}^v = 1/N_i \sum_{t=1}^{N_i} \mathbf{x}_{i,a,t}^v$ and the corresponding video motion feature $\mathbf{x}_{i,m}^v = 1/N_i \sum_{t=1}^{N_i} \mathbf{x}_{i,m,t}^v$. So the training feature set consists of N video samples can be represented by $X = \{(\mathbf{x}_{i,a}^v, \mathbf{x}_{i,m}^v, l_i)\}_{i=1}^N$, where $\mathbf{x}_{i,a}^v$ and $\mathbf{x}_{i,m}^v$ are appearance and motion feature vectors of the ith video sample labeled with l_i under vth view.

3.1. Method

Asymmetric distance metric. The video-based person re-id is to match person videos across non-overlapping camera views. As shown in Fig. 1, since both the appearance and motion of the same person could look very different at different camera views,

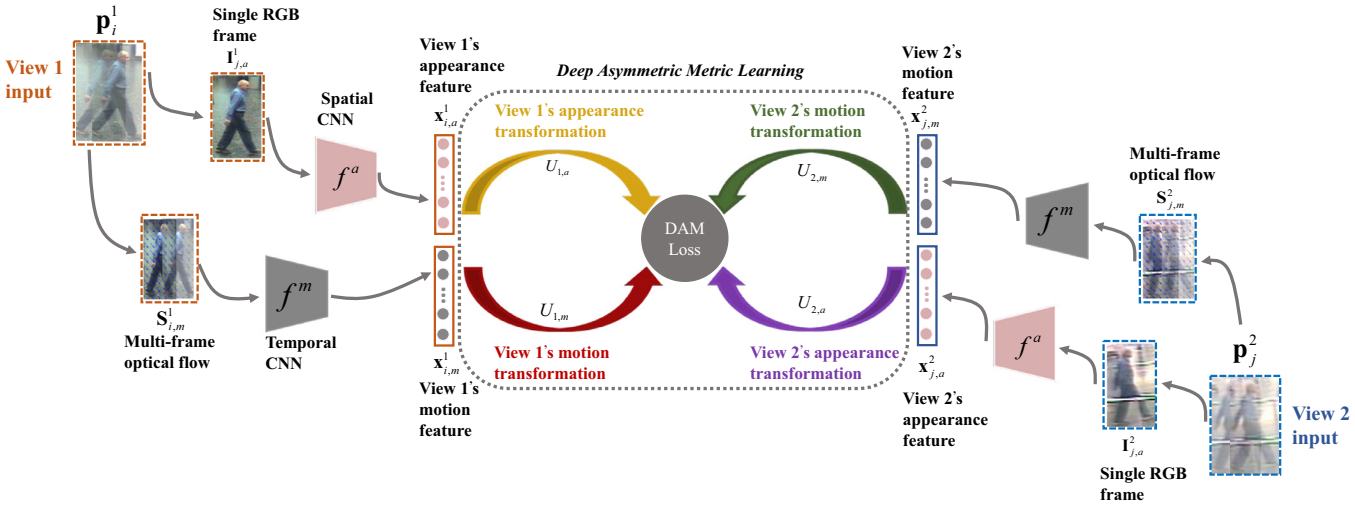


Fig. 3. Demonstration of our deep asymmetric metric learning process on two camera views network as an example. The input of the network is denoted by $\mathbf{p}_i^1$ and $\mathbf{p}_j^2$, where $\mathbf{p}_i^v = \{\mathbf{I}_{i,a}^v, \mathbf{S}_{i,m}^v\}$, the $\mathbf{I}_{i,a}^v$ represents RGB images from the v-th ($v = 1, 2, \ldots, V$) camera view and the associated $\mathbf{S}_{i,m}^v$ is the optical flow sequence which conveys the movement or gait information. Then the appearance feature $\mathbf{x}_{i,a}^v$ and motion feature $\mathbf{x}_{i,m}^v$ will be extracted using feature extractor f^a and f^m which correspond to the spatial CNN and temporal CNN, respectively. In order to alleviate the view-bias problem, the view-specific and feature-specific transformations $U_{1,a}$, $U_{1,m}$, $U_{2,a}$ and $U_{2,m}$ (where $U_{1,a} \neq U_{2,a}$, $U_{1,m} \neq U_{2,m}$) are learned by the minimization of the DAM loss.

which is called the "view-bias" problem, it could be not optimal to learn transformation for all the camera views equally. Therefore, we propose to learn the *view-specific transformation* for every camera view to alleviate this problem.

To model view-specific transformation, we learn different view-specific transformation matrices for features of different views. Let $\mathbf{x}_i^v$ denote feature vector of the ith video sample labeled with l_i under vth view. Let $U_v = [\mathbf{u}_v^1, \mathbf{u}_v^2, \ldots, \mathbf{u}_v^r]$ denote the view-specific transformation matrix under the vth view and r is the dimension of the projected subspace. These view-specific transformation matrices project the features of different views from the original space into a shared subspace as follow:

$$\mathbf{y}_i^v = U_v^T \mathbf{x}_i^v, \tag{4}$$

where the $\mathbf{y}_i^v$ is the projected feature vector of the vth view.

Moreover, it is found that people's video information can be attributed to spatial and temporal cues. As illustrated in Fig. 2, appearance feature and motion feature describe different aspects of a person and contribute unequally. And thus we also aim to learn *feature-specific transformation* for each type of feature so as to take the difference into the formulation. Overall, we form an asymmetric metric learning method to jointly learn the view-specific and feature-specific transformations to overcome the view-bias problem. Let $U_{v,a} = [\mathbf{u}_{v,a}^1, \mathbf{u}_{v,a}^2, \ldots, \mathbf{u}_{v,a}^r]$ and $U_{v,m} = [\mathbf{u}_{v,m}^1, \mathbf{u}_{v,m}^2, \ldots, \mathbf{u}_{v,m}^r]$ be the appearance and motion transformation matrices under the vth view, respectively, and r is the dimension of the projected subspace. The appearance feature and motion feature are projected from the original space into a shared subspace as follows:

$$\mathbf{y}_{i,a}^v = U_{v,a}^T \mathbf{x}_{i,a}^v, \mathbf{y}_{i,m}^v = U_{v,m}^T \mathbf{x}_{i,m}^v, \tag{5}$$

where the $\mathbf{y}_{i,a}^v$ and $\mathbf{y}_{i,m}^v$ are the projected appearance and motion feature vectors in the shared subspace at the vth view, respectively.

We consider that the appearance and motion features can be complementary to each other for video-based person re-id, the distance between two persons (e.g. $\mathbf{p}_i^v$, $\mathbf{p}_j^c$) is formed as the linear combination of appearance feature asymmetric distance and motion feature asymmetric distance, i.e.

$$d(\mathbf{p}_i^v, \mathbf{p}_j^c) = \alpha \|U_{v,a}^T \mathbf{x}_{i,a}^v - U_{c,a}^T \mathbf{x}_{j,a}^c\|_2^2 + (1-\alpha)\|U_{v,m}^T \mathbf{x}_{i,m}^v - U_{c,m}^T \mathbf{x}_{j,m}^c\|_2^2, \tag{6}$$

where $\alpha \in [0, 1]$ refers to a weighting parameter for balancing these two terms during identification.

Given the queried person, it would be challenging to find the correct matchings from the large gallery dataset. It is expected that the distance of positive pairs is smaller than the negative one. To achieve this goal, a *push term* is formed to *push the negative matching away from the positive matching*, which can be formulated as:

$$f_{push} = \sum_{\mathbf{p}_i^v, \mathbf{p}_j^c, l_i = l_j} \sum_{\mathbf{p}_i^v, \mathbf{p}_k^q, l_i \neq l_k} \max\left\{d(\mathbf{p}_i^v, \mathbf{p}_j^c) + \rho - d(\mathbf{p}_i^v, \mathbf{p}_k^q), 0\right\}, \tag{7}$$

where $d(\mathbf{p}_i^v, \mathbf{p}_j^c)$ is the intra-class distance, $d(\mathbf{p}_i^v, \mathbf{p}_k^q)$ is the inter-class distance and ρ is a parameter controls the margin between positive pairs and negative pairs.

Since there are countless positive pair and negative pair tuples, we are concerning the relative comparison between the maximum distance of positive pairs and the distance of related negative pairs, rather than comparing the negative pair with each of the relevant positive pairs, and this would notably reduce the computational complexity of our model. Thus the push term is reformed as:

$$f_{push} = \sum_{\mathbf{p}_i^v, \mathbf{p}_k^q, l_i \neq l_k} \max\left\{\max_{\mathbf{p}_i^v, \mathbf{p}_j^c, l_i = l_j} d(\mathbf{p}_i^v, \mathbf{p}_j^c) + \rho - d(\mathbf{p}_i^v, \mathbf{p}_k^q), 0\right\}. \tag{8}$$

Minimizing the *push term* is to maximize the inter-class separation, but it does not quantify the intra-class variation. Therefore, a *pull term* is formed to minimize the distance between samples of the same class in order to strengthen the correlation between samples of any positive pair, that is:

$$f_{pull} = \sum_{\mathbf{p}_i^v, \mathbf{p}_j^c, l_i = l_j} d(\mathbf{p}_i^v, \mathbf{p}_j^c). \tag{9}$$

With this term, *the positive pairs are pulled together*. As shown in the Fig. 4, the intra-class samples are moved to a cluster and close to each other after applying the pull term.

By unifying both the *push term* and *pull term*, the overall ***A**symmetric distance **M**etric learning **L**oss* (AML) for minimization is

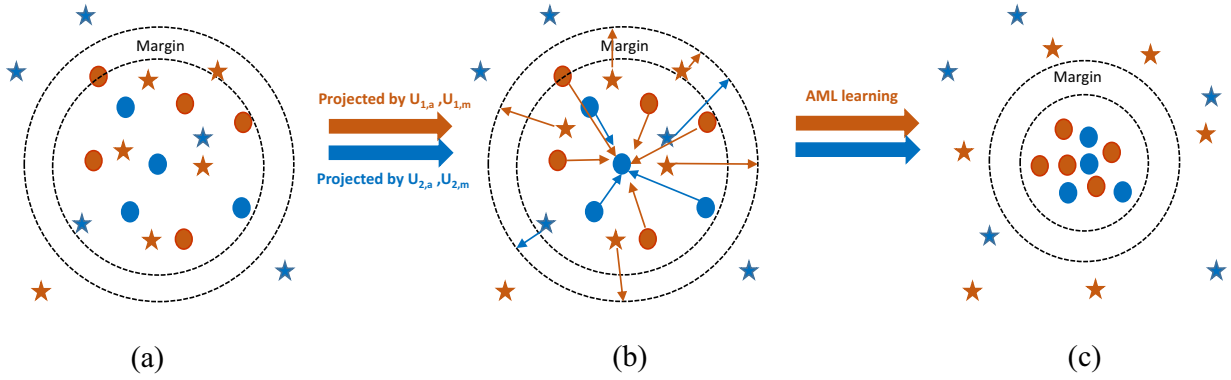


Fig. 4. A two-class toy learning process example of minimizing AML loss on two camera network as an example, where data points of the same shape are from the same class and the same color means the same camera view (red for Camera 1 and blue for Camera 2): (a) data points are randomly distributed in the original space; (b) By AML, the intra-class points are pulled together and the inter-class points are pushed away from the intra-class points by learning the view-specific and feature-specific transformations (red and blue arrows); (c) the data points are projected in a shared subspace that the distance of intra-class is smaller than inter-class. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

finally formed below:

$$\begin{aligned} f_{AML} =& f_{push} + f_{pull} + \lambda f_{reg} \\ =& \sum_{\mathbf{p}_i^v, \mathbf{p}_k^q, l_i \neq l_k} \max\left\{ \max_{\mathbf{p}_i^v, \mathbf{p}_j^c, l_i = l_j} d(\mathbf{p}_i^v, \mathbf{p}_j^c) + \rho - d(\mathbf{p}_i^v, \mathbf{p}_k^q), 0 \right\} \\ &+ \sum_{\mathbf{p}_i^v, \mathbf{p}_j^c, l_i = l_j} d(\mathbf{p}_i^v, \mathbf{p}_j^c) \\ &+ \lambda \left(\sum_{v,c,v \neq c} \left(\|U_{v,a} - U_{c,a}\|_2^2 + \|U_{v,m} - U_{c,m}\|_2^2 \right) \right. \\ &\left. + \sum_{\mathbf{p}_i^v} \sum_t \left(\|U_{v,a}^T \left(\mathbf{x}_{i,a}^v - \mathbf{x}_{i,a,t}^v \right) \|_2^2 + \|U_{v,m}^T \left(\mathbf{x}_{i,m}^v - \mathbf{x}_{i,m,t}^v \right) \|_2^2 \right) \right). \end{aligned} \tag{10}$$

In the above, the two extra terms parameterized by λ form a regularization function f_{reg} for measuring the difference between these transformations (the second last term) that is also called the camera view discrepancy regularization [21,37,86] and for reducing the intra-video variations (the last term). A toy learning process example of AML is shown in Fig. 4.

Deep embedding & network architecture. We now embed the proposed asymmetric distance metric learning loss (AML) which learns view-specific and feature-specific transformations simultaneously for acquiring heterogeneous features complementary to each other into a two-stream deep neural network. By this embedding, not only an end-to-end architecture is formed to make the feature and the asymmetric metric be learned together, but also non-linear asymmetric metric learning can be performed.

While the AML is able to enlarge the margin between classes and shrink intra-class variation as well, the classification loss is not directly quantified during training. In addition to embedding the loss function of AML, we further consider incorporating a Softmax loss function denoted as f_{CLS} in order to ensure that different videos belong to the same class are classified the same. Thus, we have our loss function for the proposed ***Deep Asymmetric Metric learning* (DAM)** method is

$$f_{DAM} = f_{AML} + f_{CLS}. \tag{11}$$

After embedding this loss into our two-stream deep neural network, an end-to-end deep neural framework DAM is built. Our DAM can jointly learn appearance feature and motion feature using two separate pipelines, and the view-specific transformation and feature-specific transformation will be learned together at the same time.

As illustrated in Fig. 3, to apply our proposed model, we adopt the two-stream architecture used in [69] to learn the appearance feature and motion feature by separate streams, where VGG-16 [87] is adopted as our basic spatial and temporal CNN network for the appearance and motion channels. Note that the parameters of spatial CNN and temporal CNN are not shared. Our AML is embedded into the network by augmenting two metric layers before the fc layer to learn the view-specific and feature-specific transformations in a non-linear way. Moreover, the margin parameter ρ and the dropout probability are set to 1 and 0.5 on all datasets, respectively. All images are resized to 224×224. The length of input optical flows is set to 10 as in [69]. The optimal values of hyper-parameters α and λ are set by cross-validation on the validation set. In the training process, we pre-train the spatial CNN sub-network and temporal CNN sub-network separately with the learning rate 0.001, combine the two stream network in a unified framework and then fine-tune the overall network with the learning rate 0.0001.

3.2. Optimization

We use the momentum method to update the adaptive weights, and the gradient back-propagation method to optimize the parameters of the network. Both of them are carried out in a mini-batch pattern. We calculate the gradients of our asymmetric distance metric learning loss f_{AML} with respect to the metric transformations (e.g. $U_{v,a}$, $U_{v,m}$, $U_{c,a}$, $U_{c,m}$, $U_{q,a}$, $U_{q,m}$). Since the process of computing gradients are fairly similar for each transformation parameter, we take the $U_{v,a}$ as an example to show the procedure of computing gradient below.

Recalling the Eq. (10), the f_{AML} is composed of f_{push}, f_{pull} and f_{reg}. For ease of understanding, we will compute the gradient of feature-specific metric $U_{v,a}$ term by term (e.g. f_{push}, f_{pull}, f_{reg}).

Firstly, for f_{push}, we find a set of indices $(i, j, k) \in \mathcal{N}$ where each index (i, j, k) triggers the push term, and then the f_{push} can be reformulated by:

$$f_{push} = \sum_{(i,j,k) \in \mathcal{N}} (d(\mathbf{p}_i^v, \mathbf{p}_j^c) + \rho - d(\mathbf{p}_i^v, \mathbf{p}_k^q)), \tag{12}$$

So the gradient can be computed by:

$$\partial f_{push} / \partial U_{v,a} = \sum_{(i,j,k) \in \mathcal{N}} \partial (d(\mathbf{p}_i^v, \mathbf{p}_j^c) - d(\mathbf{p}_i^v, \mathbf{p}_k^q)) / \partial U_{v,a}, \tag{13}$$

where the $\partial d(\mathbf{p}_i^v, \mathbf{p}_j^c)/\partial U_{v,a}$ is pre-computed by:

$$\partial d(\mathbf{p}_i^v, \mathbf{p}_j^c)/\partial U_{v,a} = 2(U_{v,a}^T \mathbf{x}_{i,a}^v - U_{c,a}^T \mathbf{x}_{j,a}^c)\mathbf{x}_{i,a}^{v\ T}. \tag{14}$$

For f_{pull}, the gradient is:

$$\partial f_{pull}/\partial U_{v,a} = \sum_{\mathbf{p}_i^v, \mathbf{p}_j^c, l_i = l_j} \partial d(\mathbf{p}_i^v, \mathbf{p}_j^c)/\partial U_{v,a}, \tag{15}$$

For f_{reg}, its gradient is calculated by:

$$\partial f_{reg}/\partial U_{v,a} = 2\lambda \left(\sum_{v,c,v\neq c} (U_{v,a} - U_{c,a}) + \sum_{\mathbf{p}_i^v} \sum_t \left(U_{v,a}^T \mathbf{x}_{i,a}^v - U_{v,a}^T \mathbf{x}_{i,a,t}^v \right) \right), \tag{16}$$

Overall, the gradient of $U_{v,a}$ is written as follow:

$$\partial f_{AML}/\partial U_{v,a} = \partial f_{push}/\partial U_{v,a} + \partial f_{pull}/\partial U_{v,a} + \partial f_{reg}/\partial U_{v,a}, \tag{17}$$

The gradients of the other transformation parameters can be computed in a similar way. In the implementation details, the symmetric layer is replaced by multiple asymmetric layers. Suppose there are totally V camera views in the camera network, and then the parameter size of the metric layer in the deep neural network is V times. When there are thousands of camera views in the camera network, it would be costly to learn view-specific transformation for each type of feature at every camera view. That is why we propose the clustering-based DAM (Section 4), for which the computation cost is the same as the DAM.

4. Clustering-based DAM

The clustering-based DAM is developed to reduce the memory cost of our DAM model when the number of camera view is too large such that learning view-specific transformation for each view costs too much memory. The general idea of the clustering-based DAM is that the camera views are clustered according to the video frames captured under that view using the agglomerative hierarchical clustering algorithm. After getting the camera view clusters, we replace learning view-specific transformation by learning cluster-specific transformation in DAM as introduced in Section 3.

The main steps of the agglomerative hierarchical clustering algorithm are: (1) assigning each data object (e.g. camera view) into its singleton group; (2) iteratively merging the two closest groups; (3) repeating the step 2) until all the data objects are merged into a certain number of (e.g. K) clusters. More specifically, given several large video-based person re-id datasets that are captured from V different camera views totally. Firstly, images (i.e. video frames) are randomly sampled from every camera view, and thus a feature subset is formed to represent for the corresponding camera view, denoted as $\mathcal{B}_v$, where $v = 1, 2, 3, \ldots, V$. Note that images' feature is extracted by a pre-trained CNN network[1] Then we initialize the singleton group by every data object $\mathcal{B}_v$ that $\mathcal{D}_v = \{\mathcal{B}_v\}$, v = 1,2,...,V.

Secondly, the distance between two singleton groups has to be calculated by measuring the difference between any two data objects $\mathcal{B}_v \in \mathbb{R}^{m\times n_v}$ and $\mathcal{B}_c \in \mathbb{R}^{m\times n_c}$ at the beginning of the clustering. For this purpose, the Wasserstein distance, which is effective in measuring the difference between data distributions [88,89], is adopted to compute the distance between two feature subsets (e.g. data points). The Wasserstein distance [90] we used to measure the difference between $\mathcal{B}_v$ and $\mathcal{B}_c$ is defined by

$$W_2(\mathcal{B}_v, \mathcal{B}_c)^2 = \frac{1}{2}\left(\|\mathbf{m}_v - \mathbf{m}_c\|_2^2 + \|\sigma_v - \sigma_c\|_2^2 \right), \tag{18}$$

[1] The pre-trained CNN network is learned using the labeled images under all camera views with Softmax loss.

where $\mathbf{m}_v, \mathbf{m}_c \in \mathbb{R}^m$ are the average of feature vectors of $\mathcal{B}_v$ and $\mathcal{B}_c$, respectively and the σ_v and σ_c are the standard deviations of $\mathcal{B}_v$ and $\mathcal{B}_c$, respectively.

Given the distance measurement between data objects, we should define the similarity of groups which contain several data objects for further clustering. The distance between two groups is represented by the average distance of all pairs of data objects belonging to different groups (e.g. $\mathcal{D}_a$, $\mathcal{D}_b$) and formally calculated by:

$$d(\mathcal{D}_a, \mathcal{D}_b) = \sum_{\mathcal{B}_v \in \mathcal{D}_a} \sum_{\mathcal{B}_c \in \mathcal{D}_b} 1/\mathcal{N}(\mathcal{D}_a) 1/\mathcal{N}(\mathcal{D}_b) W_2(\mathcal{B}_v, \mathcal{B}_c)^2, \tag{19}$$

where $\mathcal{N}(\mathcal{D}_a)$ denote the number of data objects in $\mathcal{D}_a$. Based on this equation, the two closest groups can be found and merged into one group. Iteratively repeating the merge procedure until all the data objects are merged into K groups.

After this clustering, the K groups denoted as $\mathcal{D}_k$ where $k = 1, 2, \ldots, K$ are our final clusters. Finally, we perform DAM to learn the cluster-specific transformation instead of view-specific transformation. We refer to the above procedure as the *Clustering-based DAM* algorithm, and it is summarized in Algorithm 1.

Algorithm 1: Clustering-based DAM Algorithm.

Input:
The collection of data objects: $\mathcal{B} = \{\mathcal{B}_v\}_{v=1}^V$
The number of clusters: K
Initialize the loop signal: *continue* $\leftarrow$ *true*;
Initialize every singleton group with one data object: $\mathcal{D} \leftarrow \{\mathcal{D}_v = \{\mathcal{B}_v\}\}_{v=1}^V$.
while *continue* **do**
  Search the two closest groups $\mathcal{D}_a$, $\mathcal{D}_b$ based on Eq. 19
  Update $\mathcal{D}_a \leftarrow \mathcal{D}_a \bigcup \mathcal{D}_b$
  Remove $\mathcal{D}_b$ from $\mathcal{D}$
  if $\mathcal{N}(\mathcal{D})$ *is* K **then** *continue* $\leftarrow$ *false*
end
Using DAM to learn the cluster-specific transformation
return *Cluster-specific transformation*

5. Experiments

5.1. Datasets and settings

Our experiments were conducted on three publicly available video-based person re-id datasets: PRID 2011 [91], iLIDS-VID [44] and MARS [92]. Among them, person images in PRID 2011 and iLIDS-VID were manually cropped, and person images in MARS were automatically detected.

PRID 2011. The PRID 2011 dataset consists of video pairs recorded from two different but static surveillance cameras. Totally, 385 persons are involved in camera view A, and 749 persons are in camera view B. Among all persons, 200 persons were recorded in both camera views. Each video is comprised of 5 to 675 image frames, with an average of 100 for each. This dataset was captured in uncrowded outdoor scenes with relatively simple and clean background and rare occlusions, and several different poses of person are available in each camera view (Fig. 5 (a)). The dataset was randomly divided into training set and testing set by half, and the training and testing sets both have 100 identities.

iLIDS-VID. The iLIDS-VID dataset was captured at an airport arrival hall and contains 600 video of 300 randomly sampled people. Each person has one pair of video from two camera views. Each video is comprised of 23 to 192 images frames, with an average

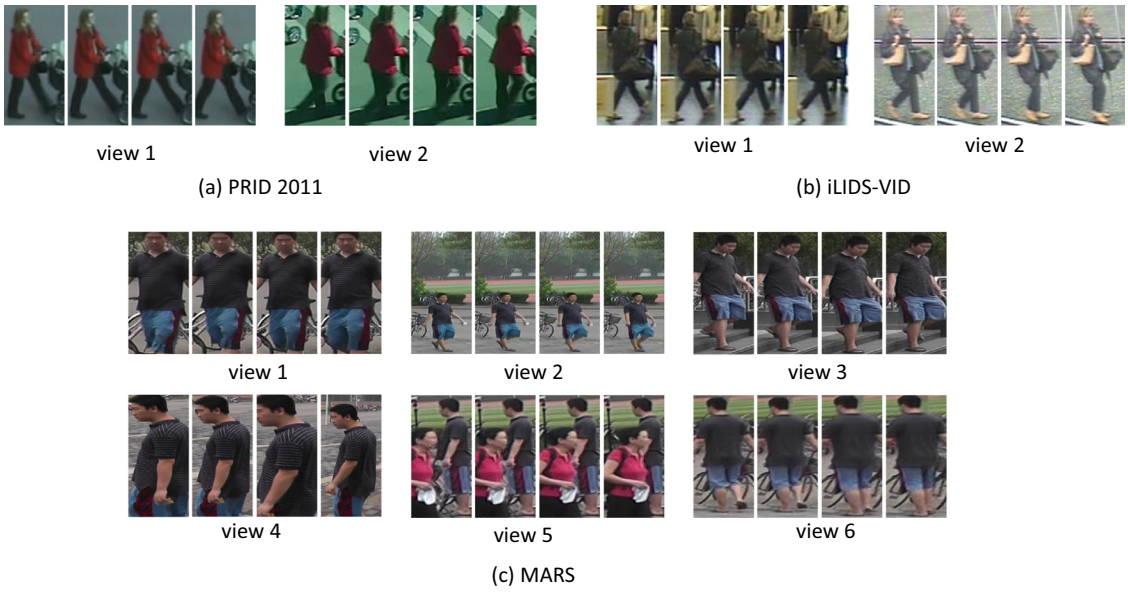


Fig. 5. Example pairs of image sequences of the same person appearing in different camera views. There are two camera views and the images of persons were manually labelled on (a) PRID 2011 and (b) iLIDS-VID dataset; on (c) MARS, there are six cameras and the tracklets were automatically generated by the DPM detector and GMMCP tracker.

of 73 for each. This dataset is more challenging due to similar clothing among people, lighting and viewpoint variations across camera views, and cluttered background with random occlusions (Fig. 5 (b)). Both the test and training set have 150 identities.

MARS. The MARS dataset was captured by six near-synchronized cameras in an university campus, containing 1261 IDs and around 20,000 tracklets where each identity has 13.2 tracklets on average. The tracklets were automatically generated by the Deformable Part Model (DPM) detector [93] and GMMCP tracker [94]. The detection/tracking error enables MARS to be more realistic. Fig. 5 (c) is the example image sequences of the same person appearing in six different camera views. Due to the inaccurate tracklets produced by false detection or tracking results, the extraction of motion feature is very challenging. The MARS dataset was divided into train and test sets, which contained 631 and 630 identities, respectively. Since each identity may have multiple tracklets under a camera view, the representative probe tracklet was randomly selected from them by following the protocol in [92].

For PRID 2011 and iLIDS-VID, the widely used Cumulative Matching Characteristics (CMC) curve is employed for quantitative measurement. For MARS, both mean Average Precision (mAP) and CMC are reported.

5.2. Evaluation of the proposed deep asymmetric metric learning (DAM)

5.2.1. Effectiveness of asymmetric modeling

In order to show the effectiveness of the asymmetric learning, we modified DAM for comparison. We produced a variant called DSM. DSM represents the symmetric version of our method that learns universal feature transformation for each type of features at all camera views. By comparing DAM with DSM in Table 1, it is clear that our asymmetric model is more effective than the symmetric one, i.e. the deep asymmetric distance learning is more effective than the deep symmetric distance metric learning. And this is because learning view-specific transformation is able to exploit visual cues specific to the environment of each camera view and thus the view-bias problem can be better alleviated.

In our asymmetric metric learning for video-based person re-id, we learn view-specific transformations for appearance and motion features, respectively. However, these two types of view-specific transformations are learned jointly rather than independently. To further evaluate this, we further compared DAM with its variant DAM_V. The difference between DAM_V and DAM is that DAM_V simply concatenates the appearance and motion features as a spatial-temporal representation in the network without learning feature-specific transformation and only learn the view-specific transformation for the concatenated one. The comparison results between DAM_V and DAM verify the usefulness of learning feature-specific transformation in our asymmetric modelling, since 2–3 matching rate is reduced at Rank-1 without learning feature-specific transformation consistently over all datasets. This is due to the fact that the contributions of different features are different, and thus quantifying them separately makes benefits. Note that sometimes simply concatenating appearance and motion features will lead to degradation as shown on the MARS dataset, and this suggests learning feature-specific transformation makes the collaboration between appearance and motion features more effective.

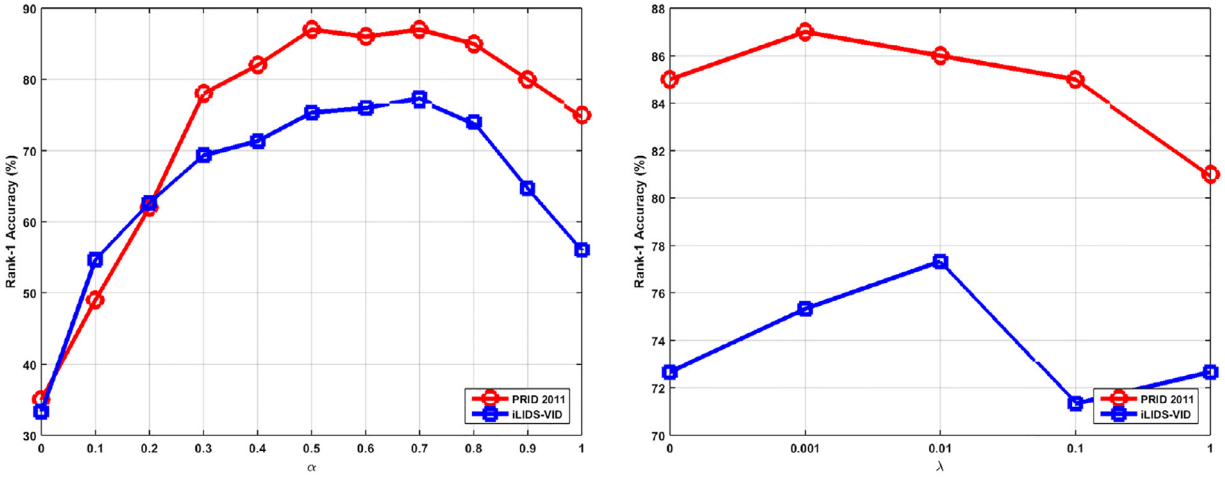
Then, we investigated the impact of parameter α, which is a hyper-parameter that balances the weight of appearance and motion information in Eq. (6) when recognizing a person. The special case when only quantifying appearance or motion features is reported in Table 2, it is true that learning both appearance and motion feature would be notably better than only learning each of them. Although appearance feature is more powerful (e.g. 56% at Rank-1 on iLIDS-VID) than motion feature (e.g. 33.33% at Rank-1 on iLIDS-VID), they are indeed complementary to each other (e.g. 77.33% at Rank-1 on iLIDS-VID by joint learning). The impact of α on the Rank-1 matching accuracy on PRID 2011 and iLIDS-VID datasets is reported in Fig. 6(a). the best performance is reached when α is around 0.7. To further understand what we have learned from DAM, we visualize the appearance and motion feature maps of the first convolution layer in Fig. 7, and it shows that the appearance feature map represents much more on the shape of a person while the motion feature map is more on the variation thing. The experiment indicates that a proper joint learning of appearance and motion features is very important to effectively use these two complementary feature transformation.

Table 1
The evaluation of primary components."DAM" is our proposed method; DSM represents the symmetric version of our method that learns universal feature transformation for each type of features at all camera views; DAM_V discards the feature-specific transformation by simply concatenating appearance and motion features as a spatial-temporal representation in the network; "WITHOUT_XX" means removing this part "XX" from our DAM. Results are shown as matching rates (%) at Rank = 1, 5, 10, 20 on all datasets and mAP on MARS dataset. The best results are in black boldface font.

Methods	PRID 2011				iLIDS-VID				MARS			
	Rank-1	Rank-5	Rank-10	Rank-20	Rank-1	Rank-5	Rank-10	Rank-20	Rank-1	Rank-5	Rank-20	mAP
DAM	**87.00**	97.00	98.00	**99.00**	**77.33**	94.00	**98.67**	**100**	**74.65**	87.02	**93.13**	**57.71**
DSM	83.00	97.00	98.00	**99.00**	73.33	93.33	94.67	96.67	71.86	88.22	92.10	55.93
DAM_V	85.00	97.00	**99.00**	**99.00**	74.66	**95.33**	98.00	**100**	72.47	**89.09**	92.36	56.38
WITHOUT_AML	75.00	93.00	97.00	**99.00**	61.33	86.00	94.00	95.33	65.77	83.11	89.53	53.44
WITHOUT_CLS	75.00	93.00	97.00	**99.00**	68.00	88.00	96.67	98.67	65.01	81.24	88.87	52.09
WITHOUT_reg	85.00	**98.00**	**99.00**	**99.00**	72.67	94.00	98.00	**100**	73.38	86.72	92.63	57.26

Table 2
The results of our method using different visual information: the "Appearance & Motion" means using the appearance and motion information at the same time, the "Appearance" ("Motion)" indicates using only appearance information (motion information). Results are shown as matching rates (%) at Rank = 1, 5, 10, 20 on all datasets and mAP on MARS dataset. The best results are in black boldface font.

Methods	PRID 2011				iLIDS-VID				MARS			
	Rank-1	Rank-5	Rank-10	Rank-20	Rank-1	Rank-5	Rank-10	Rank-20	Rank-1	Rank-5	Rank-20	mAP
Appearance & Motion	**87.00**	**97.00**	**98.00**	**99.00**	**77.33**	**94.67**	**98.67**	**100**	**74.65**	**87.02**	**93.13**	**57.71**
Appearance	75.00	94.00	97.00	98.00	56.00	84.67	94.00	97.33	73.28	86.01	93.28	57.09
Motion	35.00	66.00	81.00	93.00	33.33	68.00	75.33	87.33	10.96	23.08	38.33	4.05



(a) α (b) λ

Fig. 6. The influence of α and λ in our DAM on Rank-1 matching accuracy on PRID 2011 and iLIDS-VID datasets.

Table 3
The results of our DAM based on a similar two-stream M3D architecture. The best results are in black boldface font.

	MARS			
	Rank-1	Rank-5	Rank-20	mAP
2D CNN [70]	77.02	89.70	94.49	61.99
2D CNN [70] + DAM (our)	**79.95**	**92.17**	**95.91**	**66.51**
M3D [70]	70.20	82.98	90.25	51.95
M3D [70] + DAM (our)	**73.03**	**86.92**	**92.63**	**56.09**
Two-stream M3D [70]	77.22	89.75	94.49	62.22
Two-stream M3D [70] + DAM (our)	**80.66**	**92.27**	**96.16**	**67.38**

Finally, we embed our deep asymmetric metric learning method (DAM) into a similar two-stream network architecture [70], which is composed of a 3D CNN network stream and a 2D CNN network stream. Based on the source code provided by [70], we embed our DAM into this two-stream M3D framework and the results on MARS shown in Table 3 indicate that our DAM can be beneficial when embedded into other re-id methods, and thus it is effective and scalable.

5.2.2. Effectiveness of major components in DAM

Our DAM is composed of two primary loss functions: one for the asymmetric distance metric learning loss (AML) and the other for the classification (CLS). Also, a regularization term is considered in the formulation of DAM. In this section, we estimated the efficiency of each component by removing each of them from DAM and saw how much the performance would drop. The results were represented in Table 1, where the notation of WITHOUT_AML (WITHOUT_CLS, WITHOUT_reg) means the loss of AML (CLS, the regularization) is removed from DAM.

As reported, removing each of the components will lead to a decrease of matching performance for video-based re-id. Besides, discarding AML and CLS makes the fall notably. After eliminating AML from DAM, the matching accuracy drops strikingly, about 12, 16 and 9 matching rate on PRID 2011, iLIDS-VID and MARS datasets, respectively. This verifies the usefulness of our asymmetric modeling for alleviating the view-bias across camera

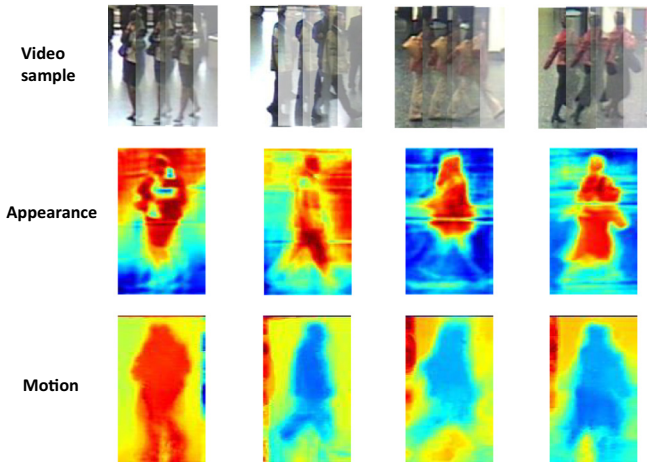


Fig. 7. The illustration of the first convolution feature map of video samples. The first row is the video samples. The bottom two rows are the corresponding appearance feature map and motion feature map, respectively, which are extracted from the first convolutional layer.

Table 4
The influence of the number of clusters on matching rates (%) by using the clustering-based DAM: the "K" is the number of the final clusters. The best results are in black boldface font.

	MARS			
	Rank-1	Rank-5	Rank-20	mAP
K=2	73.69	86.11	93.03	56.80
K=4	**74.19**	**87.42**	**93.23**	**58.26**
K=6	71.92	85.86	92.68	56.26
K=8	72.88	86.62	92.93	57.27
K=10	72.88	86.62	92.93	57.27

views during learning view-specific and as well as feature-specific transformation.

Indeed, removing CLS also makes a remarkable descent. As we have indicated, the CLS loss can make complement on directly quantifying the classification loss of the feature learning part in our DAM, which is not explicitly addressed by AML. Hence, CLS is a helpful companion for facilitating the feature learning in DAM.

Finally, we investigated the parameter λ in Fig. 6(b), the λ is used to parametrize the regularization function f_{reg}. The results point out that most best results were empirically found between $\lambda = 0.001$ and $\lambda = 0.01$, and most worst results were observed when λ is large (e.g. 5 matching rate reduced). We can also learn that the utility of regularization term is depended on a properly λ value, and when λ is large, the camera view discrepancy regularization will force our learning framework towards symmetric modeling which degrades the performance of DAM.

5.2.3. Evaluation of clustering-based DAM

Finally, we estimated the Clustering-based DAM which is used to solve the scale problem of asymmetric learning potentially. Since there is no big enough publicably available dataset that consists of lots of camera views, we combined the existing three video datasets PRID 2011, iLIDS-VID and MARS to form a mixed dataset that consists of data samples from 10 camera views. The clusters are obtained by using the agglomerative hierarchical clustering algorithm, then the cluster-specific transformation is learned by replacing view-specific transformation learning in DAM. The overall network used in this section was trained using all samples from these datasets, and the results of MARS with the different numbers of clusters K are tabulated in Table 4. Note that for person re-identification on cross camera view matching, at least two cameras exist, and thus it is natural to investigate the case when $K \geq 2$ for applying our asymmetric method. When $K = 1$, the Clustering-based DAM is equal to DSM in theory as shown in Table 1. We did not conducted Clustering-based DAM on iLIDS-VID and PRID 2011, since they only have two camera views and DAM does not have a scale problem in such a case.

On MARS, as shown, the Clustering-based DAM approximates DAM with similar matching performance, when comparing Table 4 with Table 1. Interestingly, the results suggest when K is much smaller than the number of camera views, we gained better results on MARS (e.g. 74.19% at Rank 1 when $K = 4$ by Clustering-based DAM vs. 72.88% when $K = 10$), this improvement may be a result of augmenting other datasets' data when learning the clusters of camera views and the cluster-specific transformation as well. While our experiments are on limited number of camera views due to the lack of very large public visual surveillance datasets in literatures, the report experimental results could suggest clustering the camera views and learning cluster-specific transformations instead of view-specific transformations is tractable and would not degrade the performance much, but the scale of asymmetric modelling could be reduced.

5.3. Comparison with the state-of-the-art methods

In Table 5, we reported the comparison of our method with existing state-of-the-art video-based person re-id methods including DVR [44], TDL [45], RCN [51], Deep RCN [49], SI^2DL [46], TAPR [55], BRNN [54], TAM&SRM [95], ASTPN [53], AMOC [47] Two Stream SCNN [56], QAN [65], T-CN [96] and SPW [97]. Among them, DVR [44], TDL [45], SI^2DL [46] and SPW [97] are concentrating on the distance metric learning method which using the hand-craft features such as HOG3D and STFV3D; RCN [51], Deep RCN [49], TAPR [55], BRNN [54], TAM&SRM [95], ASTPN [53], AMOC [47], two Stream SCNN [56], QAN [65] and T-CN [96] are end-to-end deep learning methods that learn feature and metric simultaneously. However, compared to our model, most of these deep methods learn equal feature transform matrix for all camera views and also do not consider the feature-specific transformation as well.

In comparison, the proposed DAM worked overall better. As we can see from the Table 5, the matching performance was improved significantly on iLIDS–VID and MARS datasets. Especially, on iLIDS-VID, the results illustrate clearly that the rank-1 matching rate of DAM was about 11% (i.e. 8 matching rate) higher as compared to the second best method SPW which indicates that our method has a better capability of handling the challenging situation which has a more complex background, and occlusion. For MARS dataset, we obtain a four matching rate improvement against the second best method TAM&SRM which contains a discriminative frame selection procedure. While these methods do not explicitly and directly model the "view-bias" problem and always extract a spatial-temporal representation without considering the specific transformation of appearance and motion features, our DAM embeds the asymmetric distance metric learning loss (AML) into a two-stream neural network to alleviate these problems and achieves overall better and stable performance. Additionally, it also demonstrates that our method has a better performance on the dataset that contains complex surveillance network which suffers from severe "view-bias" problem.

On PRID 2011 datasets, DAM performed stably as the second best with tight margin against the first one. However, DAM performed more stably, since although QAN was slightly better than ours on PRID 2011, but its performance on iLIDS-VID and MARS was clearly lower than ours. Since iLIDS-VID and MARS datasets have a more complex background and occlusion challenge relative to PRID2011 and lots of poorly cropped images exist on MARS, our results can also indicate that DAM has a better capability to handle this challenging situation.

Table 5
Comparison with the state-of-the-art video-based person re-id methods on PRID 2011, iLIDS-VID and MARS datasets. Results are shown as matching rates (%) at Rank = 1, 5, 10, 20 on all datasets and mAP on MARS dataset. The best results are in black boldface font.

Methods	PRID 2011				iLIDS-VID				MARS			
	Rank-1	Rank-5	Rank-10	Rank-20	Rank-1	Rank-5	Rank-10	Rank-20	Rank-1	Rank-5	Rank-20	mAP
DAM	87.00	97.00	98.00	99.00	**77.33**	94.00	**98.67**	**100**	**74.65**	87.02	93.13	**57.71**
DVR [44]	40.0	71.7	84.5	92.2	39.5	61.1	71.7	81.0	–	–	–	–
TDL [45]	56.74	80.00	87.64	93.59	56.33	87.60	95.60	98.27	–	–	–	–
RCN [51]	70	90	95	97	58	84	91	96	–	–	–	–
Deep RCN [49]	69.0	88.4	93.2	96.4	46.1	76.8	89.7	95.6	–	–	–	–
SI^2DL [46]	76.7	95.6	96.7	98.9	48.7	81.1	89.2	97.3	–	–	–	–
TAPR [55]	73.9	94.6	97.4	98.9	55.0	87.5	93.8	97.2	–	–	–	–
BRNN [54]	72.8	92.0	95.1	97.6	55.3	85.0	91.7	95.1	–	–	–	–
TAM&SRM [95]	79.4	94.4	–	99.3	55.2	86.5	–	97.0	70.6	**90.0**	**97.6**	50.7
ASTPN [53]	77	95	99	99	62	86	94	98	44	70	81	–
AMOC [47]	83.7	**98.3**	**99.4**	**100**	68.7	**94.3**	98.3	99.3	68.3	81.4	90.6	**52.9**
Two Stream SCNN [56]	78	94	97	99	60	86	93	97	–	–	–	–
QAN [65]	**90.3**	98.2	99.32	100.0	68.0	86.8	95.4	97.4	57.37	71.16	79.80	32.27
T-CN [96]	81.1	95.0	97.3	98.7	60.6	83.8	91.2	95.8	–	–	–	–
SPW [97]	83.5	96.3	98.5	**100.0**	69.3	89.6	95.7	98.2	–	–	–	–

6. Conclusion

We have addressed the "view-bias" problem in video-based person re-id, and we have developed a novel end-to-end deep asymmetric metric learning (DAM) for solving this problem. Compared to existing video-based person re-id, DAM is a deep asymmetric distance metric learning and enables learning view-specific and feature-specific transformations jointly, so that it is able to learn specific transformations for different features at different camera views and thus better solves the "view-bias" problem. DAM has been further extended to a clustering-based DAM to make it scalable in a large-scale camera network. Extensive evaluations have been carried out on three public datasets iLIDS-VID, PRID2011 and MARS for verifying our proposed DAM. A limitation of our method is that it is still not efficient enough for online video processing since the motion feature in our framework relies on optical flow, which is widely used for modelling motion and requires considerable computation cost. In the future, we would like to seek more efficient motion feature for developing a more efficient video-based re-id system.

Acknowledgment

This work was supported partially by the National Key Research and Development Program of China (2016YFB1001002), NSFC(61522115,U1811461), Guangdong Province Science and Technology Innovation Leading Talents (2016TX03X157), and the Royal Society Newton Advanced Fellowship (NA150459).

References

[1] M. Farenzena, L. Bazzani, A. Perina, V. Murino, M. Cristani, Person re-identification by symmetry-driven accumulation of local features, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2010, pp. 2360–2367.
[2] B. Ma, Y. Su, F. Jurie, Local descriptors encoded by fisher vectors for person re-identification, in: Proceedings of the European Conference on Computer Vision Workshops and Demonstrations, 2012, pp. 413–422.
[3] R. Zhao, W. Ouyang, X. Wang, Unsupervised salience learning for person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2013, pp. 3586–3593.
[4] S. Liao, Y. Hu, X. Zhu, S.Z. Li, Person re-identification by local maximal occurrence representation and metric learning, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2015, pp. 2197–2206.
[5] T. Matsukawa, T. Okabe, E. Suzuki, Y. Sato, Hierarchical gaussian descriptor for person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2016, pp. 1363–1372.
[6] C. Zhao, X. Wang, D. Miao, H. Wang, W. Zheng, Y. Xu, D. Zhang, Maximal granularity structure and generalized multi-view discriminant analysis for person re-identification, Pattern Recognit. 79 (2018) 79–96.
[7] B. Prosser, W.-S. Zheng, S. Gong, T. Xiang, Q. Mary, Person re-identification by support vector ranking, in: Proceedings of the British Machine Vision Conference, 2010, p. 6.
[8] W.-S. Zheng, S. Gong, T. Xiang, Person re-identification by probabilistic relative distance comparison, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2011, pp. 649–656.
[9] M. Hirzer, P.M. Roth, M. Köstinger, H. Bischof, Relaxed pairwise learned metric for person re-identification, in: Proceedings of the European Conference on Computer Vision, Springer, 2012, pp. 780–793.
[10] M. Koestinger, M. Hirzer, P. Wohlhart, P.M. Roth, H. Bischof, Large scale metric learning from equivalence constraints, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2012, pp. 2288–2295.
[11] S. Pedagadi, J. Orwell, S. Velastin, B. Boghossian, Local fisher discriminant analysis for pedestrian re-identification, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2013, pp. 3318–3325.
[12] R. Zhao, W. Ouyang, X. Wang, Person re-identification by salience matching, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 2528–2535.
[13] W.-S. Zheng, S. Gong, T. Xiang, Reidentification by relative distance comparison, IEEE Trans. Pattern Anal. Mach. Intell. 35 (3) (2013) 653–668.
[14] K. Fukunaga, Introduction to Statistical Pattern Recognition, Academic Press, 2013.
[15] F. Xiong, M. Gou, O. Camps, M. Sznaier, Person re-identification using kernel-based metric learning methods, in: Proceedings of the European Conference on Computer Vision, Springer, 2014, pp. 1–16.
[16] G. Lisanti, I. Masi, A.D. Bagdanov, A. Del Bimbo, Person re-identification by iterative re-weighted sparse ranking, IEEE Trans. Pattern Anal. Mach. Intell. 37 (8) (2015) 1629–1642.
[17] N. Li, R. Jin, Z.-H. Zhou, Top rank optimization in linear time, in: Proceedings of the Advances in Neural Information Processing Systems, 2014, pp. 1502–1510.
[18] S. Paisitkriangkrai, C. Shen, A. van den Hengel, Learning to rank in person re-identification with metric ensembles, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2015, pp. 1846–1855.
[19] G. Lisanti, I. Masi, A.D. Bagdanov, A. Del Bimbo, Person re-identification by iterative re-weighted sparse ranking, IEEE Trans. Pattern Anal. Mach. Intell. (2015) 1629–1642.
[20] Y.-C. Chen, W.-S. Zheng, J. Lai, Mirror representation for modeling view-specific transform in person re-identification, in: Proceedings of the International Joint Conference on Artificial Intelligence, 2015, pp. 3402–3408.
[21] Y.-C. Chen, W.-S. Zheng, J.-H. Lai, P. Yuen, An asymmetric distance model for cross-view feature mapping in person re-identification, IEEE Trans. Circuits Syst. Video Technol. (2016).
[22] L. Zhang, T. Xiang, S. Gong, Learning a discriminative null space for person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2016, pp. 1239–1248.
[23] S. Bak, P. Carr, One-shot metric learning for person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2017.
[24] Z. Zhong, L. Zheng, D. Cao, S. Li, Re-ranking person re-identification with k-reciprocal encoding, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2017, pp. 3652–3661.
[25] W. Chen, X. Chen, J. Zhang, K. Huang, Beyond triplet loss: a deep quadruplet network for person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2, 2017.
[26] S. Zhou, J. Wang, J. Wang, Y. Gong, N. Zheng, Point to set similarity based deep feature learning for person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 6, 2017.
[27] J. Chen, Y. Wang, J. Qin, L. Liu, L. Shao, Fast person re-identification via cross-camera semantic binary transformation, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2017.

[28] Y. Ren, X. Li, X. Lu, Feedback mechanism based iterative metric learning for person re-identification, Pattern Recognit. 75 (2018) 99–111.
[29] J. Wang, Z. Wang, C. Liang, C. Gao, N. Sang, Equidistance constrained metric learning for person re-identification, Pattern Recognit. 74 (2018) 38–51.
[30] J. Dai, Y. Zhang, H. Lu, H. Wang, Cross-view semantic projection learning for person re-identification, Pattern Recognit. 75 (2018) 63–76.
[31] J. Li, A.J. Ma, P.C. Yuen, Semi-supervised region metric learning for person re-identification, Int. J. Comput. Vis. 126 (8) (2018) 855–874.
[32] W.-S. Zheng, S. Gong, T. Xiang, Towards open-world person re-identification by one-shot group-based verification, IEEE Trans. Pattern Anal. Mach. Intell. 38 (3) (2016) 591–606.
[33] W. Li, R. Zhao, T. Xiao, X. Wang, Deepreid: deep filter pairing neural network for person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2014, pp. 152–159.
[34] E. Ahmed, M. Jones, T.K. Marks, An improved deep learning architecture for person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2015, pp. 3908–3916.
[35] S. Ding, L. Lin, G. Wang, H. Chao, Deep feature learning with relative distance comparison for person re-identification, Pattern Recognit. 48 (10) (2015) 2993–3003.
[36] T. Xiao, H. Li, W. Ouyang, X. Wang, Learning deep feature representations with domain guided dropout for person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2016, pp. 1249–1258.
[37] Y.-C. Chen, X. Zhu, W.-S. Zheng, J.-H. Lai, Person re-identification by camera correlation aware feature augmentation, IEEE Trans. Pattern Anal. Mach. Intell. 40 (2) (2018) 392–408.
[38] D. Li, X. Chen, Z. Zhang, K. Huang, Learning deep context-aware features over body and latent parts for person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2017, pp. 384–393.
[39] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, X. Tang, Spindle net: person re-identification with human body region guided feature decomposition and fusion, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2017, pp. 1077–1085.
[40] J. Lin, L. Ren, J. Lu, J. Feng, J. Zhou, Consistent-aware deep learning for person re-identification in a camera network, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 6, 2017.
[41] S. Zhou, J. Wang, D. Meng, X. Xin, Y. Li, Y. Gong, N. Zheng, Deep self-paced learning for person re-identification, Pattern Recognit. 76 (2018) 739–751.
[42] L. Wu, Y. Wang, J. Gao, X. Li, Deep adaptive feature embedding with local sample distributions for person re-identification, Pattern Recognit. 73 (2018) 275–288.
[43] D. Cheng, Y. Gong, X. Chang, W. Shi, A. Hauptmann, N. Zheng, Deep feature learning via structured graph Laplacian embedding for person re-identification, Pattern Recognit. 82 (2018) 94–104.
[44] T. Wang, S. Gong, X. Zhu, S. Wang, Person re-identification by discriminative selection in video ranking, IEEE Trans. Pattern Anal. Mach. Intell. 38 (12) (2016) 2501–2514.
[45] J. You, A. Wu, X. Li, W.-S. Zheng, Top-push video-based person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2016, pp. 1345–1353.
[46] X. Zhu, X.-Y. Jing, F. Wu, H. Feng, Video-based person re-identification by simultaneously learning intra-video and inter-video distance metrics, in: Proceedings of the International Joint Conference on Artificial Intelligence, 2016, pp. 3552–3559.
[47] H. Liu, Z. Jie, K. Jayashree, M. Qi, J. Jiang, S. Yan, J. Feng, Video-based person re-identification with accumulative motion context, IEEE Trans. Circuits Syst. Video Technol. 28 (10) (2018) 2788–2802.
[48] L. Wu, C. Shen, A. van den Hengel, Convolutional LSTM networks for video-based person re-identification. arXiv:1606.01609.
[49] L. Wu, C. Shen, A.v.d. Hengel, Deep recurrent convolutional networks for video-based person re-identification: an end-to-end approach. arXiv:1606.01609.
[50] W. Zhang, S. Hu, K. Liu, Learning compact appearance representation for video-based person re-identification. arXiv:1702.06294.
[51] N. McLaughlin, J. Martinez del Rincon, P. Miller, Recurrent convolutional network for video-based person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2016, pp. 1325–1334.
[52] Z. Zhou, Y. Huang, W. Wang, L. Wang, T. Tan, See the forest for the trees: joint spatial and temporal recurrent neural networks for video-based person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2017, pp. 6776–6785.
[53] S. Xu, Y. Cheng, K. Gu, Y. Yang, S. Chang, P. Zhou, Jointly attentive spatial-temporal pooling networks for video-based person re-identification. arXiv:1708.02286.
[54] W. Zhang, X. Yu, X. He, Learning bidirectional temporal cues for video-based person re-identification, IEEE Trans. Circuits Syst. Video Technol. 28 (10) (2017) 2768–2776.
[55] C. Gao, J. Wang, L. Liu, J.-G. Yu, N. Sang, Temporally aligned pooling representation for video-based person re-identification, in: Proceedings of the IEEE International Conference on Image Processing, 2016, pp. 4284–4288.
[56] D. Chung, K. Tahboub, E.J. Delp, A two stream siamese convolutional neural network for person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2017, pp. 1983–1991.
[57] C. Su, S. Zhang, F. Yang, G. Zhang, Q. Tian, W. Gao, L.S. Davis, Attributes driven tracklet-to-tracklet person re-identification using latent prototypes space mapping, Pattern Recognit. 66 (2017) 4–15.
[58] D. Simonnet, M. Lewandowski, S. Velastin, J. Orwell, E. Turkbeyler, Re-identification of pedestrians in crowds using dynamic time warping, in: Proceedings of the European Conference on Computer Vision Workshops and Demonstrations, 2012, pp. 423–432.
[59] X. Lan, S. Zhang, P.C. Yuen, R. Chellappa, Learning common and feature-specific patterns: a novel multiple-sparse-representation-based tracker, IEEE Trans. Image Process. 27 (4) (2018) 2022–2037.
[60] S. Zhang, Y. Qi, F. Jiang, X. Lan, P.C. Yuen, H. Zhou, Point-to-set distance metric learning on deep representations for visual tracking, IEEE Trans. Intell. Transp. Syst. 19 (1) (2018) 187–198.
[61] K. Bozek, L. Hebert, A.S. Mikheyev, G.J. Stephens, Towards dense object tracking in a 2D honeybee hive, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4185–4193.
[62] A. Klaser, M. Marszałek, C. Schmid, A spatio-temporal descriptor based on 3D–gradients, in: Proceedings of the British Machine Vision Conference, British Machine Vision Association, 2008. 275–1
[63] K. Liu, B. Ma, W. Zhang, R. Huang, A spatio-temporal appearance representation for viceo-based pedestrian re-identification, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 3810–3818.
[64] W.Z.H.L. Yiheng Liu Zhenxun Yuan, Spatial and temporal mutual promotion for video-based person re-identification, in: Proceedings of the Association for the Advancement of Artificial Intelligence, 2019.
[65] Y. Liu, J. Yan, W. Ouyang, Quality aware network for set to set recognition, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2017, pp. 5790–5799.
[66] S. Li, S. Bak, P. Carr, X. Wang, Diversity regularized spatiotemporal attention for video-based person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 369–378.
[67] D. Chen, H. Li, T. Xiao, S. Yi, X. Wang, 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, 2018, pp. 1169–1178.
[68] Y. Fu, X. Wang, Y. Wei, T. Huang, STA: spatial-temporal attention for large-scale video-based person re-identification, in: Proceedings of the Association for the Advancement of Artificial Intelligence, 2019.
[69] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, in: Proceedings of the Advances in Neural Information Processing Systems, 2014, pp. 568–576.
[70] J. Li, S. Zhang, T. Huang, Multi-scale 3D convolution network for video based person re-identification, in: Proceedings of the Association for the Advancement of Artificial Intelligence, 2019.
[71] C. Xu, D. Tao, C. Xu, A survey on multi-view learning. arXiv:1304.5634.
[72] C. Xu, D. Tao, C. Xu, Multi-view learning with incomplete views, IEEE Trans. Image Process. 24 (12) (2015) 5812–5825.
[73] J. Li, C. Xu, W. Yang, C. Sun, D. Tao, Discriminative multi-view interactive image re-ranking, IEEE Trans. Image Process. 26 (7) (2017) 3113–3127.
[74] G. Lisanti, S. Karaman, I. Masi, Multichannel-kernel canonical correlation analysis for cross-view person reidentification, ACM Trans. Multimed. Comput. Commun. Appl. 13 (2) (2017) 13.
[75] L. An, S. Yang, B. Bhanu, Person re-identification by robust canonical correlation analysis, IEEE Signal Process. Lett. 22 (8) (2015) 1103–1107.
[76] J. Hu, J. Lu, Y.-P. Tan, Discriminative deep metric learning for face verification in the wild, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2014, pp. 1875–1882.
[77] J. Yu, X. Yang, F. Gao, D. Tao, Deep multimodal distance metric learning using click constraints for image ranking, IEEE Trans. Cybern. 47 (12) (2017) 4014–4024.
[78] H. Oh Song, Y. Xiang, S. Jegelka, S. Savarese, Deep metric learning via lifted structured feature embedding, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2016, pp. 4004–4012.
[79] J. Lu, G. Wang, W. Deng, P. Moulin, J. Zhou, Multi-manifold deep metric learning for image set classification, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2015, pp. 1137–1145.
[80] J. Hu, J. Lu, Y.-P. Tan, Deep transfer metric learning, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2015, pp. 325–333.
[81] Y. Cui, F. Zhou, Y. Lin, S. Belongie, Fine-grained categorization and dataset bootstrapping using deep metric learning with humans in the loop, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2016, pp. 1153–1162.
[82] X. Han, T. Leung, Y. Jia, R. Sukthankar, A.C. Berg, MatchNet: unifying feature and metric learning for patch-based matching, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2015, pp. 3279–3286.
[83] D. Yi, Z. Lei, S. Liao, S.Z. Li, Deep metric learning for person re-identification, in: Proceedings of the International Conference on Pattern Recognition, 2014, pp. 34–39.
[84] E. Hoffer, N. Ailon, Deep metric learning using triplet network, in: Proceedings of the International Workshop on Similarity-Based Pattern Recognition, 2015, pp. 84–92.
[85] Y.-C. Chen, X. Zhu, W.-S. Zheng, J.-H. Lai, Person re-identification by camera correlation aware feature augmentation, IEEE Trans. Pattern Anal. Mach. Intell. 40 (2) (2018) 392–408.

[86] X. Zhu, B. Wu, D. Huang, W.-S. Zheng, Fast open-world person re-identification, IEEE Trans. Image Process. (2017).
[87] K. Simonyan, A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556.
[88] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein gan. arXiv:1701.07875.
[89] D. Berthelot, T. Schumm, L. Metz, Began: boundary equilibrium generative adversarial networks. arXiv:1703.10717.
[90] R. He, X. Wu, Z. Sun, T. Tan, Wasserstein CNN: learning invariant features for NIR-VIS face recognition. arXiv:1708.02412.
[91] M. Hirzer, C. Beleznai, P.M. Roth, H. Bischof, Person re-identification by descriptive and discriminative classification, in: Proceedings of the Scandinavian Conference on Image Analysis, 2011.
[92] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, Q. Tian, Mars: a video benchmark for large-scale person re-identification, in: Proceedings of the European Conference on Computer Vision, 2016, pp. 868–884.
[93] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, IEEE Trans. Pattern Anal. Mach. Intell. 32 (9) (2010) 1627–1645.
[94] A. Dehghan, S. Modiri Assari, M. Shah, GMMCP tracker: globally optimal generalized maximum multi clique problem for multiple object tracking, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2015, pp. 4091–4099.
[95] Z. Zhou, Y. Huang, W. Wang, L. Wang, T. Tan, See the forest for the trees: joint spatial and temporal recurrent neural networks for video-based person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2017, pp. 6776–6785.
[96] Y. Wu, J. Qiu, T. Jun, O. Tsukasa, Temporal-enhanced convolutional network for person re-identification, in: Proceedings of the Association for the Advancement of Artificial Intelligence, 2018.
[97] W. Huang, C. Liang, Y. Yu, Z. Wang, W. Ruan, R. Hu, Video-based person re-identification via self paced weighting, in: Proceedings of the Association for the Advancement of Artificial Intelligence, 2018.

Jingke Meng received the bachelor's degree in computer science and technology from Sun Yat-Sen University in 2015. She is pursuing Ph.D. degree with the School of Data and Computer Science in Sun Yat-sen University. Hers research interests are computer vision and person re-identification.

Ancong Wu received the bachelor's degree in intelligence science and technology from Sun Yat-Sen University in 2015. He is pursuing Ph.D. degree with the School of Electronics and Information Technology in Sun Yat-sen University. His research interests are computer vision and person re-identification. URL: http://isee.sysu.edu.cn/~wuancong.

Wei-Shi Zheng is now a Professor with Sun Yat-sen University. He received the Ph.D. degree in Applied Mathematics from Sun Yat-sen University in 2008. He is now a full Professor at Sun Yat-sen University. He was a postdoctoral researcher on the EU FP7 SAMURAI Project at Queen Mary University of London and an associate professor at Sun Yat-sen University after that. He has now published more than 100 papers, including more than 70 publications in major journals (TPAMI,TNN/TNNLS,TIP,TSMC-B,PR) and top conferences (ICCV, CVPR,IJCAI,AAAI). He has joined the organisation of four tutorial presentations in ACCV 2012, ICPR 2012, ICCV 2013 and CVPR 2015. His research interests include person/object association and activity understanding in visual surveillance, and the related large-scale machine learning algorithm. Especially, He has active research on person re-identification in the last five years. He serves a lot for many journals and conference, and he was announced to perform outstanding review in recent top conferences (ECCV 2016 & CVPR 2017). He has ever joined Microsoft Research Asia Young Faculty Visiting Programme. He has ever served as an area chair/associate editor of AVSS 2012, ICPR 2018, and BMVC 2018. He is an associate editor of Pattern Recognition. He is a recipient of Excellent Young Scientists Fund of the National Natural Science Foundation of China, and a recipient of Royal Society-Newton Advanced Fellowship of United Kingdom. URL: http://isee.sysu.edu.cn/~zhwshi/.