This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2017.2748698, IEEE Transactions on Circuits and Systems for Video Technology

P2SNet : Can an Image Match a Video for Person Re-identification in an End-to-end Way?

Guangcong Wang, Jianhuang Lai*, *Senior Member, IEEE,* and Xiaohua Xie, *Member, IEEE*

***Abstract*—We address a new person re-identification problem, in which our goal is to directly match image-based scenarios with video-based ones. This differs significantly from the conventional person re-identification problem, which aims to match two image-based scenarios (and it is assumed that the available video frames have been manually selected to form the image-based scenarios). To solve this more challenging and realistic problem without the implicit assumption of manual selection, we propose an end-to-end matching framework called a point-to-set network (P2SNet), which consists of 1) a *k*-nearest neighbor triplet (*k*NN-triplet) module, which functions as a "denoiser" by letting the network sequentially focus on the available frames while ignoring the other useless frames in a video; and 2) a novel deep neural network that uses videos and images as input to jointly learn the feature representations and a point-to-set distance metric in a unified way. Our P2SNet is evaluated on three new image-to-video person re-identification datasets, i-LIDS-VID-P2S, PRID2011-P2S and MARS-P2S, which are modified from i-LIDS-VID, PRID 2011 and MARS, respectively. The experimental results demonstrate the superior performance of our model over the other state-of-the-art methods.**

***Index Terms*—person re-identification, image-to-video matching, *k*NN-triplet, deep learning.**

I. Introduction

PERSON re-identification (re-id), which aims to match pedestrian images across multiple non-overlapped cameras, has been attracting increasing research attention because it has many applications in video surveillance for public security and safety. To match a person across views, the similarity between pairs of pedestrian data taken from different data sources needs to be measured. Usually, the data sources include closed-circuit television (CCTV) photos and videos. Thus, it is essential to obtain an appropriate discriminant similarity measure for the matching types of pedestrian data, such as image-to-image (image-based), video-to-video (video-based), and image-to-video. Among these metrics, the image-to-video similarity is more challenging to measure due to the heterogeneous nature of the data. However, image-to-video pedestrian matching is more commonplace than image-to-image matching, because video surveillance systems tend to only capture videos. Image-to-video matching is sometimes used in urgent situations, for example, when a criminal is being chased, the police often have to examine surveillance videos from all around the area to match a photo taken by someone who last saw the criminal (see Fig. 1). The goal is to find the person with the same appearance in videos taken by other cameras given only a probe image. We call this the ***image-to-video re-identification*** problem.

There are three computational challenges in solving the image-to-video person re-id problem (see the examples in Fig. 1). First, the pedestrian data from different sources are heterogeneous. Specifically, an image can be regarded as a point in a high dimensional space, whereas a video is treated as a set. It is extremely difficult to define an appropriate point-to-set distance metric for the image-to-video person re-id problem. Second, given a large number of surveillance videos, it is difficult to extract effective features from the videos given the noise in real-world scenarios. For example, in the Market-1501 dataset [1], the pedestrians are first cropped by the pedestrian detector DPM [2] and then manually selected to discard "distractors" and "junk". However, the manual selection of pedestrians is impractical in real-world scenarios. Because the conventional approaches extract the features from the video set without using de-noising algorithms, unavailable frames (e.g., false alarm detections and pedestrians with occlusions) are also considered in the system. Third, it is especially difficult to jointly learn the effective features and appropriate point-to-set distance metric in an end-to-end unified framework.

The existing person re-id approaches cannot solve the image-to-video person re-id problem in an end-to-end manner, regardless of whether they are designed for computing cross-view invariant features or distance metrics [3]–[5]. The existing models can be divided into two groups: image-to-image matching [5], [25], [33] and video-to-video [6]–[9]. However, both groups share the method of sampling and average pooling without removing the unavailable images/frames. With the video-to-video method, some frames are selected from both videos, the distances between each pair of frames from the different videos are then calculated, and the average value over all of the distances is determined without removing the outliers. This problem is also reflected in the existing benchmarking person re-id datasets, most of which are designed for either image-to-image matching or video-to-video matching. None of the datasets are designed for image-to-video matching, which uses images to form a probe set given a gallery of videos.

This project was supported by the National Natural Science Foundation of China (U1611461, 61573387, 61672544). (*Corresponding author: Jianhuang Lai.)

Guangcong Wang is with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, Guangdong 510006, China and with Guangdong Key Laboratory of Information Security Technology. E-mail: wanggc3@mail2.sysu.edu.cn.

Jianhuang Lai is with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, Guangdong 510006, China and with School of Information Science and Technology, XinHua College, Sun Yat-sen University. E-mail: stsljh@mail.sysu.edu.cn

Xiaohua Xie is with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, Guangdong 510006, China and with Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China. E-mail: xiexiaoh6@mail.sysu.edu.cn.

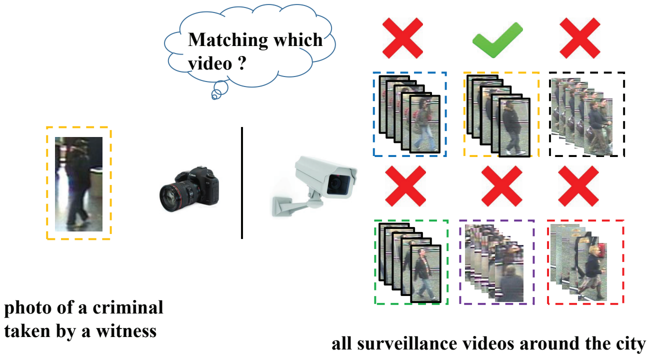

Matching which video ?
photo of a criminal taken by a witness
all surveillance videos around the city

Fig. 1. An illustration of the image-to-video person re-id problem. Here, a photo was taken by the citizen who last saw the criminal. The police seek to determine which surveillance videos the criminal appears in by matching the photo to the videos. The video with a green tick indicates the ground truth.

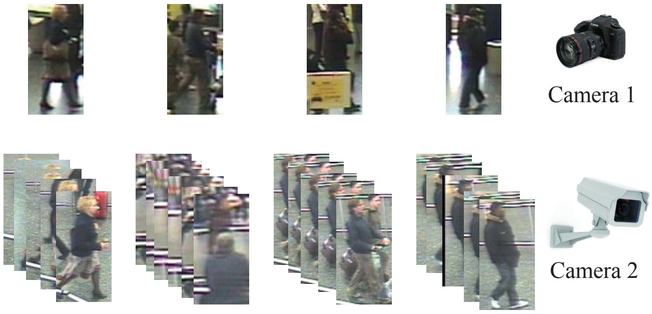


Fig. 2. Examples of image-to-video person re-id datasets. The first row indicates the person images taken by Camera 1, while the second row denotes the person video taken by Camera 2. Note that the data from different sources (i.e., image and video) are not only heterogeneous, but also have significantly different lightings, which makes the image-to-video person re-id problem very challenging.

To address the image-to-video person re-id problem in an end-to-end way, we formulate a new person re- id framework. First, we build a novel deep neural network using both videos and images as input to jointly learn the feature representations and a point-to-set distance metric in a unified way. In the literature, most of the models are executed in a pre-defined feature space; that is, the feature extraction and similarity measure are studied separately. Our architecture takes raw data from different sources as input and automatically produces their representations according to the proposed network. The feature learning and similarity metric learning are thus integrated for end-to-end optimization. Second, we use a *k*NN-triplet unit to separate the related frames from the noisy videos without any manual selection, as illustrated in Fig. 4.

In summary, this paper makes two main contributions to the literature on the person re-id problem.

- First, we integrate feature learning and point-to-set distance metric learning by building an end-to-end deep architecture of a neural network. A *k*NN-triplet module lets the network sequentially focus on the important frames while ignoring the other useless frames in a video. Consequently, our system can automatically find and select the available frames from a video.
- Second, we conduct extensive experiments to evaluate our model. In particular, because the existing benchmarks largely ignore the image-to-video problem, we modify three existing datasets to simulate the image-to-video problem (see Fig. 2). Our experimental results on three datasets show that the proposed model is effective in addressing the image-to-video problem, and achieves significant improvement over the image-to-image methods applied to the same problem. Our model also outperforms a number of alternative models designed for other point-to-set problems, such as the less challenging image-to-video face recognition.

The rest of the paper is organized as follows. Section II reviews the related work. Section III introduces our P2SNet model. Section IV presents the proposed deep neural network architecture and the learning algorithm. The experimental results, comparisons, ablation studies, and analyses are presented in Section V. Section VI concludes the paper.

II. Related Work

Person re-id aims to measure the similarity of pairs of pedestrian data. The majority of the existing methods include two separate steps: robust feature extraction and metric learning. The first step is to design a feature representation that is sufficiently robust to the lighting, poses, and viewpoint changes. For example, several features such as the color, texture, and shape are included in person re-id [10]–[12]. Second, many of the traditional methods seek to determine a variation insensitive and discriminant similarity measure [13]–[15] by learning a robust distance or similarity function to solve the complex matching problem. However, this pipeline of person re-id research has a number of potential problems. For example, extracting hand-crafted features and then learning the metrics may lead to sub-optimal solutions. Moreover, all of these works either explicitly or implicitly match the same types of image-based (video-based) pedestrian data. Thus, they are ineffective for solving the image-to-video person re-id problem, as is shown in our experiments.

In some person re-id studies, a video-based model is considered that offers a partial solution to the image-to-video problem. Wang et al. [9] presented a model that automatically selected the most discriminative video fragments from noisy image sequences of people in which more reliable space-time features can be extracted, while simultaneously learning a video ranking function for person re-id. You et al. [16] proposed a top- push distance learning model, which integrated a top-push constraint for matching the video features of persons. However, these video-to-video person re-id methods have limited success in resolving the image-to-video re-id problem in real-world scenarios. First, they usually select some representative frames from both videos, calculate the distance between each pair of frames from the different videos, and then output the average value over all of the distances. Consequently, the temporal average methods degrade the results because they do not use de-noising algorithms and also may lose intra-class information because of the smoothing of different viewpoints and poses. Second, the video-based matching models and image-based models are homogeneous, and therefore some features from the video-based models are unavailable for the image-to-video scenarios due to the heterogeneous matching problem, e.g., the HOG features extracted from the images do not match the HOG3D features extracted from the videos.

The image-to-video matching problem has also been studied extensively in other computer vision problems, and is an especially important topic in face recognition. Wang *et al.* [17] defined a point-to-manifold distance for matching points against sets. Zhou *et al.* [18] introduced a robust estimate of the posterior distribution of the identity variable to address the point-to-set issue. Huang *et al.* [19] explored the real-world still-to-video (S2V) face recognition scenario and then developed a coupling alignments with recognition (CAR) method to tightly couple the quality alignment, geometric alignment, and face recognition via low-rank regularized sparse representation in a unified framework. However, the image-to-video person re-id problem has two unique characteristics that distinguish it from the still-to-video face recognition problem addressed in these methods. First, the image-to-video problem in person re-id is much more unstructured because a persons body appearance (e.g., clothing), free poses, and deformation are much more diverse than those of the face (see Fig. 2). Second, a pedestrian video has richer information than a face video because the crowed public space is filled with the background clutter caused by other people in the scene or static obstacles such as walls/pillars. Consequently, the models of point-to-set face recognition are not effective enough to extend to the challenging person re-id task.

To overcome these unique challenges in person re-id, we propose an end-to-end matching framework called the point-to-set network (P2SNet). This network integrates feature learning and point-to-set distance metric learning by building an end-to-end deep neural network architecture. A *k*NN-triplet module enables the network to sequentially focus on the available frames while ignoring the other useless frames in a video. Our framework is motivated by a number of recently proposed deep learning methods and metric learning models, namely, convolutional neural networks (CNNs), the triplet scheme [5], and large margin nearest neighbor classification (LMNN) [20]. CNNs have attracted significant attention in visual recognition, largely due to their powerful performance [21]–[23], [32]. To jointly optimize the robust features and metric learning, several deep learning models have been proposed [3]–[5] that are based on a scalable distance driven feature learning framework and an effective triplet or pairwise generation scheme. However, the models assume that the datasets have been manually selected and thus have no noise. Nonetheless, the model proposed in [5] can be regarded as a global distance metric. For each iteration, a fixed number of classes (persons) and a fixed number of images for each class are randomly selected. Then, large distances between inputs with the same label and small distances between inputs with different labels are penalized. After many iterations, the algorithm forces all of the examples in the same class to be clustered and repels imposters away from the perimeter. However, the performance of the global distance metric may be harmed, particularly when the classes exhibit multimodal data distributions, and the data contain a significant amount of noise. To address this problem, we integrate a local distance metric (*k*NN-triplet) into the network.

III. P2SNet Model

In Section III-A, we firstly present a brief introduction of triplet loss [5] for the image-to-image matching. We then generalize this approach to the image-to-video matching in Section III-B. Furthermore, we integrate a simple but effective proximity-based outlier detection algorithm [24] into the triplet loss layer to remove outliers.

A. Triplet Loss

Given two sets of images $I = \{I_1, I_2, ..., I_N\}$ and $I^{'} = \{I^{'}_1, I^{'}_2, ..., I^{'}_N\}$, where N denotes the total number of pedestrians, a pair $(I_i, I^{'}_i)$ $(1 \leq i \leq N)$ indicates two images for the ith person from non-overlapping camera views.

First, we define the triplet set as $\Upsilon =< (I_i, I^{'}_i, I^{'}_j) >$, where $1 \leq i, j \leq N$, $i \neq j$, i.e., $(I_i, I^{'}_i)$ belongs to the same pedestrian, while $(I_i, I^{'}_j)$ belongs to different pedestrians. The model aims to minimize the distance between pair $(I_i, I^{'}_i)$ and maximize the distance between pair $(I_i, I^{'}_j)$. In a standard triplet loss network, the hinge loss is used

$$\begin{aligned} L = \sum_{\forall (I_i, I^{'}_i, I^{'}_j) \in \Upsilon} & max((d(\varphi(\mathbf{x}_i), \phi(\mathbf{x}^{'}_i)) + \alpha \\ & -d(\varphi(\mathbf{x}_i), \phi(\mathbf{x}^{'}_j))), 0), \end{aligned} \tag{1}$$

where $d(.,.)$ is the Euclidean distance between two feature vectors and α is a predefined constant parameter representing the minimum margin between matched and mismatched pairs. Two images from non-overlapping camera views share the same function extraction function $\mathbf{x}_i = f(I_i)$ and $\mathbf{x}^{'}_i = f(I^{'}_i)$. After that, the two functions $\varphi(.)$ and $\phi(.)$ enable the alignment of the two feature distributions across disjoint views for the same person. Let $\mathbf{y}_i = \varphi(\mathbf{x}_i)$ and $\mathbf{y}^{'}_i = \phi(\mathbf{x}^{'}_i)$ as aligned vectors, Eq. (1) is simplified as follows

$$L = \sum_{\forall (I_i, I^{'}_i, I^{'}_j) \in \Upsilon} max(((\mathbf{y}_i - \mathbf{y}^{'}_i)^2 + \alpha - (\mathbf{y}_i - \mathbf{y}^{'}_j)^2), 0), \tag{2}$$

let $l = (\mathbf{y}_i - \mathbf{y}^{'}_i)^2 + \alpha - (\mathbf{y}_i - \mathbf{y}^{'}_j)^2$, $\forall (I_i, I^{'}_i, I^{'}_j) \in \Upsilon$, we compute the partial derivative of three vectors by

$$\begin{cases} \dfrac{\partial l}{\partial \mathbf{y}_i} = 2(\mathbf{y}_j - \mathbf{y}^{'}_i) \\ \dfrac{\partial l}{\partial \mathbf{y}_j} = 2(\mathbf{y}_j - \mathbf{y}_i) \quad \text{if } l > 0 \\ \dfrac{\partial l}{\partial \mathbf{y}^{'}_i} = 2(\mathbf{y}_i - \mathbf{y}^{'}_i) \end{cases} \tag{3}$$

According to Eq. (3), we can collect the sum of gradient for each image and then perform the back-propagation. The forward-propagation and back-propagation of the triplet loss layer can be implemented by Eq. (2) and Eq. (3), respectively. Consequently, we can learn a similarity metric between two images from non-overlapping camera views.

Sub-network	Structure
	Input: $300 \times 3 \times 230 \times 80$ RGB
Image/frame feature	Conv-ReLU $(32, 5 \times 5, 2 \times 2, 0 \times 0)$
	Max-Pooling $(3 \times 3, 3 \times 3, 0 \times 0)$
	Conv-ReLU $(32, 5 \times 5, 1 \times 1, 0 \times 0)$
	Max-Pooling $(3 \times 3, 3 \times 3, 0 \times 0)$
	FC-ReLU (400)
Domain alignment	Slice $(slice\_dim : 0, slice\_point : 50)$
	FC *two pathways* $\times$ (400)
	Concat $axis : 0$
	Norm (L2)
Similarity measure	kNN-TripletLoss

TABLE I
THE SPECIFICATIONS OF THE PROPOSED NETWORK.

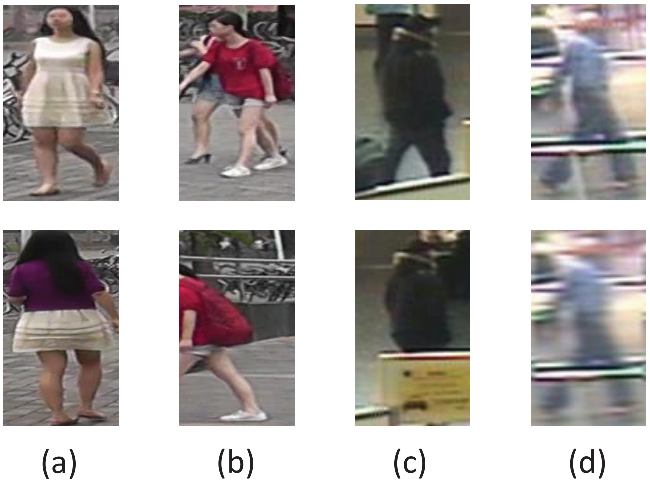


Fig. 3. Four kinds of outliers: (a) false alarm, (b) partial body, (c) occlusion, and (d) motion blur. Note that, (a) and (b) are from the MARS dataset while (c) and (d) are from the iLIDS-VID dataset. The images in the first row are the ground truth. The images in the second row are the corresponding outliers.

B. kNN-triplet Loss

Given a collection of images $I = \{I_1, I_2, ..., I_N\}$ and a collection of videos $V = \{V_1, V_2, ..., V_N\}$, where N denotes the total number of pedestrians, a pair (I_i, V_i) $(1 \leq i \leq N)$ indicates an image and a video for the ith person in a database. For each video V_i, $V_i = \{\tilde{I}_{i1}, \tilde{I}_{i2}, ..., \tilde{I}_{iL_i}\}$, where L_i denotes the number of frames in video V_i, and $\tilde{I}_{ip}$ represents the pth frame in video V_i. For the image-to-video matching, we generalize the triplet set as $\Upsilon =< (I_i, \tilde{I}_{ip}, \tilde{I}_{jq}) >$, where $1 \leq i, j \leq N$, $i \neq j$, $\forall p \leq L_i$, $\forall q \leq L_j$, i.e., $(I_i, \tilde{I}_{ip})$ belongs to the same pedestrian, while $(I_i, \tilde{I}_{jq})$ belongs to different pedestrians. Given the generalized triplet, we can easily obtain the loss according to Eq. (2)

$$L = \sum_{\forall (I_i, \tilde{I}_{ip}, \tilde{I}_{jq}) \in \Upsilon} max(((\mathbf{y}_i - \tilde{\mathbf{y}}_{ip})^2 + \alpha - (\mathbf{y}_i - \tilde{\mathbf{y}}_{jq})^2), 0), \tag{4}$$

where $\mathbf{y}_i$ and $\tilde{\mathbf{y}}_{ip}$ denote two aligned vectors of I_i and $\tilde{I}_{ip}$, respectively.

However, the global distance metric often misses the important non-linear structures in the data and is not sufficiently robust enough to the noise that results from the imperfect data collection process. For example, pedestrian detectors and trackers will lead to false alarm, partial body, part occlusion, and motion blur in real-world scenarios, as shown in Fig. 3. This imperfect data significantly does affect the performance of most of classifiers. We thus call it outlier in this paper. To address this problem, we introduce a simple but effective kNN-triplet method. For each image, we select k nearest neighbor frames from a video clip according to the outlier score for each pedestrian. The kNN-triplet set Υ^* is given by

$$\Upsilon^* =< I_i, \tilde{I}^*_{ip}, \tilde{I}^*_{jq} >, \tag{5}$$

where $\tilde{I}^*_{ip}$ and $\tilde{I}^*_{jq}$ denote the k-nearest neighbors of the image I_i, I_j, respectively. However, we can not directly compute the distance between an image and the frames with the same label, because the image may be closer to the noisy frames than the ones available in the initial training phase. Instead, we integrate a simple but effective proximity-based outlier detection algorithm [24] into the triplet loss layer for video frames. The outlier score of $\tilde{I}_{ip}$ is given by distance to the k-nearest neighbor:

$$\tilde{s}_{ip} = \frac{1}{L_i} \sum_{q=1}^{L_i} (\tilde{\mathbf{y}}_{ip} - \tilde{\mathbf{y}}_{iq})^2 \tag{6}$$

where $\tilde{\mathbf{y}}_{ip} = \phi(\tilde{\mathbf{x}}_{ip})$ is an aligned frame feature vector. A point with higher score is more likely to be an outlier. Suppose the function $Rank(.)$ sorts the elements in ascending order and outputs the ranking of each element. Given a video clip $\{\tilde{I}_{ip}\}, p = 1, 2, ..., L_i$, we define a k-nearest neighbor by

$$\tilde{I}^*_{ip} \in \{\tilde{I}_{ip'} | \{\tilde{r}_{ip}\} = Rank(\{\tilde{s}_{ip}\}), \tilde{r}_{ip'} \leq k\} \tag{7}$$

Thus we rewrite the objective function as

$$L = \sum_{\forall (I_i, \tilde{I}_{ip}, \tilde{I}_{jq}) \in \Upsilon^*} max(((\mathbf{y}_i - \tilde{\mathbf{y}}_{ip})^2 + \alpha - (\mathbf{y}_i - \tilde{\mathbf{y}}_{jq})^2), 0), \tag{8}$$

The first term in Eq. (8) represents the intra-class distance of pair (I_i, V_i), while the second term represents the inter-class distance of pair (I_i, V_j). Our model aims to minimize the distance between pair (I_i, V_i) and maximize the distance between pair (I_i, V_j). Note that we do not need to directly optimize Eq. (8). Instead, we use several convolutional layers to approximate the function $f(.)$. We also use two fully connected layers to approximate the function $\varphi(.)$ and $\phi(.)$, respectively. Let $l = (\mathbf{y}_i - \tilde{\mathbf{y}}_{ip})^2 + \alpha - (\mathbf{y}_i - \tilde{\mathbf{y}}_{jq})^2$, $\forall (I_i, \tilde{I}_{ip}, \tilde{I}_{jq}) \in \Upsilon^*$, we compute the partial derivative of three vectors by

$$\begin{cases} \dfrac{\partial l}{\partial \mathbf{y}_i} = 2(\tilde{\mathbf{y}}_{ip} - \tilde{\mathbf{y}}_{jq}) \\ \dfrac{\partial l}{\partial \tilde{\mathbf{y}}_{ip}} = 2(\tilde{\mathbf{y}}_{ip} - \mathbf{y}_i) \quad \text{if } l > 0 \\ \dfrac{\partial l}{\partial \tilde{\mathbf{y}}_{jq}} = 2(\mathbf{y}_i - \tilde{\mathbf{y}}_{jq}) \end{cases} \tag{9}$$

According to Eq. (9), we can collect the sum of gradient for each image and then perform the back-propagation. The forward-propagation and back-propagation of the kNN-triplet loss layer can be implemented by Eq. (8) and Eq. (9), respectively. Consequently, we can learn a similarity metric between an image and a video from non-overlapping camera views. See Fig. 4 and Algorithm 1 for the detail.

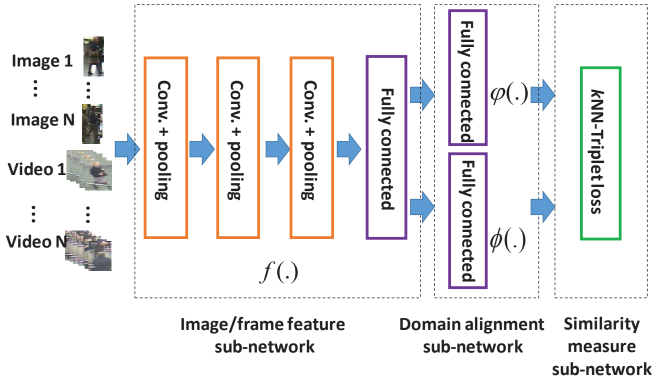


Fig. 4. Deep architecture of our P2SNet model. The architecture is comprised of three parts: the image/frame feature sub-network, domain alignment sub-network, and similarity measure sub-network. The image/frame feature sub-network extracts the feature representation from samples of images and video frames, which are built upon a number of convolutional layers, max-pooling operations, and fully-connected layers. The domain alignment sub-network aims to align the feature representations from two domains. The similarity measure sub-network includes a kNN-triplet layer that incorporates a kNN local distance metric to automatically remove the outliers.

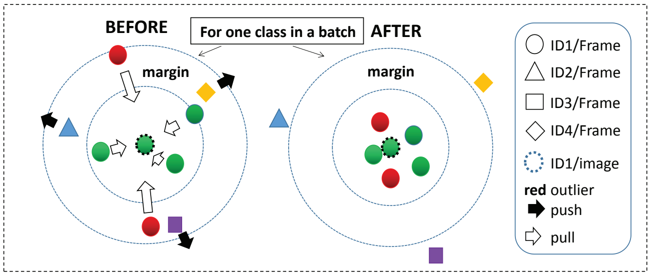


Fig. 5. Schematic illustration of the triplet for each batch before training (left) versus after training (right). The distance metric is optimized so that: (i) the samples with the same ID are pulled into a small radius, and (ii) the samples with different IDs are pushed out of this small radius by some finite margin. However, the outliers (red circles) with the same ID are also pulled into the small radius.

IV. Joint Point-to-Set Distance Metric and Feature Learning

In this section, we introduce our deep architecture that integrates point-to-set distance metric learning with convolutional feature representation learning.

A. Deep Architecture

As mentioned, our model can jointly handle point-to-set distance metric learning and feature learning. This integration is achieved by building a deep architecture of convolutional neural networks, which is illustrated in Fig. 4. This architecture is comprised of three parts: the image/frame feature sub-network, domain alignment sub-network and similarity measure sub-network. The specifications of the proposed network are shown in Table I. Note that, convolutional layers, pooling layers and fully connected layers are denoted by *(feature maps, kernel, stride, pad)*, *(kernel, stride, pad)*, and *(channel)*, respectively.

Image/frame feature sub-network. To determine whether an image and a video are of the same person, we first need to extract the image and frame features. The images and the corresponding videos are separately stored in the memory one by one as shown in Fig. 4. The first three layers of our network are convolutional layers (including three max pooling layers), which are used to compute the convolutional features. Then the network is followed by a fully connected layer, which is used to reduce the dimensionality. Thus, we approximate the function $f(.)$.

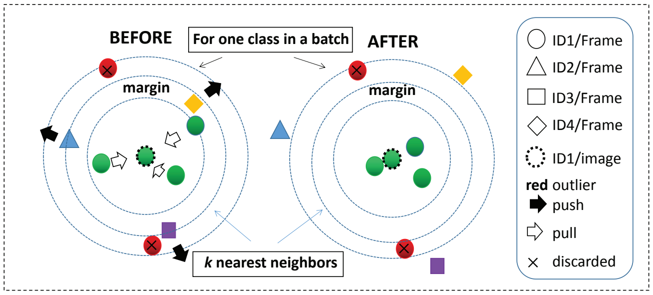


Fig. 6. Schematic illustration of the kNN-triplet for each batch before training (left) versus after training (right). Different from the conventional triplet, the kNN-triplet is introduced to remove the outliers. For each image, we compute the k nearest neighbors determined by the proximity-based outlier detection algorithm. Then the k nearest neighbors are considered to contribute to the gradient while the $T - k$ samples with the same ID (outliers) are discarded.

Domain alignment sub-network. Next, the network is sliced into two pathways, with one for images and the other for videos. Due to the changes of viewpoints, lighting conditions and camera features, images of the same person from different views differ in appearance, and thus the feature representations across the disjoint camera views follow different distributions. Domain alignment aims to align the feature representations from different domains [25]. Here, we use φ and ϕ as two linear functions to align the features such that different domains following different distributions are mapped to a common domain.

Similarity measure sub-network. Finally, following the pioneering work of [5], we apply a modified triplet loss layer to the top layer such that the true matched pairs are closer than the mismatched pairs. Different from [5], we adopt a kNN-triplet scheme. In the absence of prior knowledge, the simplest prescription is to compute the k nearest neighbors with the same class label, as determined by Eq. (6) in each batch. The differences between the triplet and kNN-triplet are shown in Fig. 5 and 6. For a conventional triplet, the samples with the same label are pulled into a small radius while those with different labels are pushed out of this small radius by some finite margin. In this way, the outliers (red circles) with the same label are also pulled into the small radius, which may harm the performance. To solve this problem, a kNN local distance metric is proposed to remove the outliers by a simple proximity-based outlier detection algorithm in the training phase. For each image, we compute the k nearest neighbors determined by the proximity-based outlier detection algorithm. Then, the k nearest neighbors are considered to contribute to the gradient while the $T - k$ samples with the same ID (outliers) are discarded. Note that in one batch, T frames are randomly sampled for each video. The details of the model training are presented in Section IV-B.

B. Model Training

In this section, we discuss the learning method for our P2SNet model. To avoid loading all of the images into the memory, we use a mini-batch learning approach, whereby, in each training iteration, a subset of image-to-video pairs is fed into the neural network for model optimization.

Our proposed framework aims to learn features to maximize the inter-class relative distance and minimize the intra-class relative distance. A novel triplet generation scheme based on an image-based gradient decent algorithm was presented in [5]. Based on the image-based gradient descent algorithm, we develop an image-video-based gradient descent algorithm combined with *k*NN decisions. Specifically, in our setting, each triplet contains one image and two frames, such as $(I_i, \tilde{I}_{ip}, \tilde{I}_{jq})$. First, we randomly select B image-to-video pairs from N image-video pairs as a batch and then feed them into the network. Note that T frames are randomly sampled for each video. At the top of the network, we get B image feature vectors and $B \times T$ frame vectors. Second, for each image feature vector, we compute its k nearest neighbor frames with the same identity, and therefore B image vectors and $B \times k$ frame vectors are considered to generate the triplets while $B \times (T - k)$ frame vectors are discarded ($k \leq T$). Third, we randomly generate G triplets per image-video pair in the loss layer, compute the loss, and then backward propagate. The details of our algorithm are shown in **Algorithm 1** below.

Algorithm 1: Image-video-based gradient descent algorithm

```
Input: $B$ aligned image vectors $\mathbf{y}_i = \varphi(\mathbf{x}_i)$, $B \times T$
       aligned frame vectors $\tilde{\mathbf{y}}_{ip} = \phi(\tilde{\mathbf{x}}_{ip})$, $\forall i \leq B$,
       $\forall p \leq T$
Output: The network parameters
for $i = 1 : B$ do
    The outlier score of $\tilde{I}_{ip}$ is given by the distance
    to its k-nearest neighbor: Compute the outlier
    score of $\tilde{\mathbf{y}}_{ip}$, $\tilde{s}_{ip} = \frac{1}{T} \sum_{q=1}^{T} d(\tilde{\mathbf{y}}_{ip}, \tilde{\mathbf{y}}_{iq})$;
    Get the ranking $\{\tilde{r}_{ip}\}$ by Rank($\{\tilde{s}_{ip}\}$);
    Keep $k$ nearest neighbor frames $\{\tilde{I}^*_{ip}\}$ according
    to $\{\tilde{r}_{ip}\}$, and discard the others;
end
for $i = 1 : B$ do
    We randomly generate G kNN-triplets
    $(I_i, \tilde{I}^*_{ip}, \tilde{I}^*_{jq})$;
end
Get a kNN-triplet set $\Upsilon^* =< I_i, \tilde{I}^*_{ip}, \tilde{I}^*_{jq} >$;
foreach $(I_i, \tilde{I}^*_{ip}, \tilde{I}^*_{jq}) \in \Upsilon^*$ do
    $l = (\mathbf{y}_i - \tilde{\mathbf{y}}_{ip})^2 - (\mathbf{y}_i - \tilde{\mathbf{y}}_{jq})^2$;
    if $l \leq -1.0$ then
        Continue;
    end
    $\frac{\partial l}{\partial \mathbf{y}_i} + = 2(\mathbf{y}_{jq} - \mathbf{y}_{ip})$; $\frac{\partial l}{\partial \mathbf{y}_{ip}} + = 2(\mathbf{y}_{ip} - \mathbf{y}_i)$;
    $\frac{\partial l}{\partial \mathbf{y}_{jq}} + = 2(\mathbf{y}_i - \mathbf{y}_{jq})$;
end
Update the network parameters.
```

V. Experiments

In this section, we apply our P2SNet model to three benchmark datasets for evaluation, i.e., iLIDS-VID-P2S, PRID2011-P2S and MARS-P2S. In these tasks, we compare the state-of-the-art methods with our model. We also present ablation studies to reveal the benefits of each main component of our method, e.g., the point-to-set distance measure and the joint optimization of the CNN feature representation and metric learning.

Datasets. No image-to-video person re-identification dataset is publicly available. Thus, we adjust the PRID 2011 [26], iLIDS-VID [9] and MARS [27] datasets to form the iLIDS-VID-P2S, PRID2011-P2S and MARS-P2S datasets, respectively, where "P2S" denotes "point to set".

Both PRID 2011 and iLIDS-VID datasets were originally developed for image-to-image and video-to-video re-id tasks. Each dataset comprises images from two cameras. The images from each camera are static based and sequence based. For our adjustment, we select static images from one camera and image sequences from another. Specifically, the iLIDS- VID dataset comprises observations of 300 pedestrians from two distinct camera views in public open space. Compared with the iLIDS-VID dataset, the PRID 2011 contains different numbers of persons, with one camera view showing 385 and the other showing 749 persons. The first 200 persons appear in both camera views. In our experiments, we only use the first 200 persons, which enables us obtain the two new datasets, iLIDS-VID-P2S and PRID2011-P2S, for examining the proposed image-to-video person re-id problem.

MARS is the largest video-based re-id benchmark dataset. It contains 1,261 identities and around 20,000 video sequences. The dataset is divided into training and test sets, containing 631 and 630 identities, respectively. Each identity has 13.2 sequences on average. Each sequence is collected from one of six different cameras. There are 2,009 probes selected for query which are the first frames of videos. In addition, the dataset also contains 3,248 distractor sequences.

Experimental setting. We use mini-batch learning in our experiments to reduce the memory requirements. In each task, we randomly select a batch of samples from the original training set to generate a number of triplets. The initial parameters of the convolutional and full connection layers are set by two zero-mean Gaussian distributions, whose standard deviations are 0.01 and 0.001, respectively. Following [27], for MARS-P2S, we use the CNN feature learned on its training data; for iLIDS-VID-P2S and PRID2011-P2S, we first pretrain the CNN models on MARS-P2S and then fine-tune on iLID-LIDS-P2S and PRID2011-P2S, respectively. The other specific settings for the different tasks are included in the following sub-sections.

Evaluation Metric. To evaluate these three benchmarks,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2017.2748698, IEEE Transactions on Circuits and Systems for Video Technology

the testing set is further divided into a gallery set of videos (i.e., one video per person) and a probe set (including images of individuals from different camera views in contrast to the gallery set). We use the cumulative matching characteristic (CMC) [28] as the evaluation metric in this task.

Data Augmentation. In our model training, all of the images are resized to 250×100, and cropped to 230×80 at the center with a small random perturbation. In every learning round, 1,200 pairs of samples are constructed by selecting 20 persons (or classes) and constructing 60 pairs for each person (class).

A. Evaluations on the iLIDS-VID-P2S Dataset

The iLIDS-VID-P2S dataset, which was created based on the iLIDS-VID dataset contains 300 images and 300 videos of 300 pedestrians collected at an airport arrival hall under a multi-camera CCTV network. Each person is observed from two distinct cameras, with an image taken by one camera and a video by the other. This dataset is very challenging due to the clothing similarities among the people, the lighting and viewpoint variations across the camera views, the cluttered background, and the occlusions (Fig. 2). We randomly partition the dataset 10 times, and obtain a training set (of 150 persons) and a testing set (of 150 persons) without overlap. We compare our approach with several state-of-the-art methods, which can be grouped into four categories. First, we use three point-to-set (P2S) manifold learning methods [29] that were originally designed for face recognition, namely, Euclidean metric learning (denoted as "LERM-ES1"), Euclidean-to-Euclidean metric learning (LERM-ES2) and Euclidean-to-Riemannian metric learning (LERM-ES3). Different from their original implementation, we use the state-of-the-art handcrafted of feature local maximal occurrence representation (denoted as "LOMO") [30] instead of the handcrafted feature of histogram equalization [29]. In Fig. 7, they are denoted as "hand+LERM-ES1", "hand+LERM-ES1", and "hand+LERM-ES1", respectively.

Second, four representative point-to-point (P2P) re-id models are used in the experiment: the KISSME distance learning method [31], MAHAL, L2 and XQDA [30]. However, these person re-id methods were not designed to solve the image-to-video person re-id, and thus not expected to be competitive. Hence, we also use a sampling method that addresses the related image-to-video face recognition problem; i.e., we do nothing with the probe images, and randomly select one frame from each gallery video as its representative. Thus, the distance is calculated between the image and the frame representative. For the point-to-point re-id, deep features [27] are used to represent an image of a person. Note that, for fair comparisons, we implement the model [27] with the same network (see Fig. 4) for feature extraction. In Fig. 7, they are denoted as "deep+XQDA (T=1)", "deep+KISSME (T=1)", "deep+MAHAL (T=1)", and "deep+L2 (T=1)", respectively.

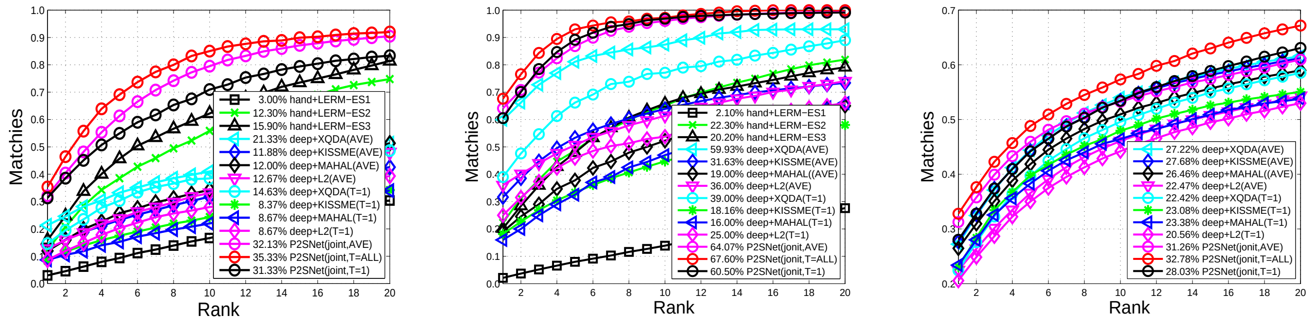
Third, we perform average pooling on the frames and then do the point-to-center matching using the above four representative point-to-point re-id models. In Fig. 7, they are denoted as "deep+XQDA (AVE)", "deep+KISSME (AVE)", "deep+MAHAL (AVE)", and "deep+L2 (AVE)", respectively.

Fourth, we also implement two end-to-end baselines for comparisons. Different from P2SNet (joint, T=ALL), P2SNet (joint, T=1) only uses one frame for test (using the same training model). P2SNet (joint, AVE) integrates the average temporal pooling into the network for training. Specially, it computes the center of all frames and learns a metric between an image and the corresponding center point by the triplet scheme [5].

The results are reported in Fig. 7(a). It is encouraging to see that our approach significantly outperforms the competing methods, e.g., by improving the state-of-the-art rank-1 accuracy from 15.90% to 35.33% compared with the manifold learning models, from 14.63% to 35.33% compared with the point-to-point models, from 21.33% to 35.33% compared with "point-to-center" models, and from 32.13% to 35.33% compared with two end-to-end baselines. Among the competing methods, although "hand+LERM-ES1", "hand+LERM-ES2", and "hand+LERM-ES3" are state-of-the-art point-to-set models used in facial recognition, they do not work well in the person re-id task, mostly because of the free poses, deformation, and background clutter. "Deep+XQDA (AVE)", "deep+KISSME (AVE)", "deep+MAHAL (AVE)", and "deep+L2 (AVE)" use deep features for point-to-set matching by computing the center of all frames, but not an end-to-end manner, which may harm the performance. Alternatively, "deep+XQDA (T=1)", "deep+KISSME (T=1)", "deep+MAHAL (T=1)", and "deep+L2 (T=1)" use a sampling method that randomly selects one frame from a video to perform the point-to-point matching. We can see that the result of the point-to-point models drops significantly, which can be attributed to the appropriateness of the point-to-set similarity for resolving the image-to-video person re-id problem. Our approach is also superior to the P2Net (joint, AVE) method that simply integrates the average temporal pooling into the network for training. This is because the average temporal pooling method may lose intra-class information, e.g., the smoothing of different viewpoints and poses. Another comparison of these two methods includes the capability of removing outliers, as discussed in Section V-D.

B. Evaluations on the PRID2011-P2S Dataset

The PRID2011-P2S dataset, which was created based on the PRID 2011 database, contains an image and a video of 200 individuals taken from distinct cameras. We randomly partition this dataset into a training set and a testing set, with 100 individuals used for testing and the others for training. Compared with the iLIDS-VID-P2S dataset, the PRID2011-P2S dataset was captured in uncrowded outdoor scenes with a relative simple and clean background and few occlusions. The competitive methods we use on the PRID2011-P2S dataset are the same as those used on the iLIDS-VID-P2S dataset. Fig. 7(b) shows the results of our method and the other competing approaches on the PRID2011-P2S dataset. According to the quantitative results, our method improves the state-of-the-art rank-1 accuracy from 22.30% to 67.60% compared with the manifold learning models, from 59.93% to 67.60% compared with the deep feature models (both point-to-center and point-



(a) iLIDS-VID-P2S

(b) PRID2011-P2S

(c) MARS-P2S

Fig. 7. CMC curves on (a) iLIDS-VID-P2S dataset; (b) PRID2011-P2S dataset (c) MARS-P2S dataset. Our method has superior performances over existing state-of-the-art methods.

to-point models), and from 64.07% to 67.60% compared with two end-to-end baselines.

C. Evaluations on the MARS Dataset

The MARS-P2S dataset, which was created based on the MARS database, contains 1,261 identities taken from distinct cameras. Following [27], in our image-to-video setting, the "image" is chosen as the first frame of a video. The dataset is divided into training and test sets, containing 631 and 630 identities, respectively. Note that, each identity may contain some distractor tracklets due to false detection or tracking results.

The competitive methods we use on the MARS-P2S dataset are proposed in [27]. For fair comparison, we implement the same network for feature extraction. Fig. 7(c) shows the results of our method and the other competing approaches on MARS-P2S. According to the quantitative results, our method improves the state-of-the-art rank-1 accuracy from 27.68% to 32.78% compared with the deep feature models, and from 31.26% to 32.78% compared with two end-to-end baselines. In this dataset, the manifold learning models are not used to compare because they need a large amount of space and time to run on the MARS dataset, e.g., it contains a very big matrix (d, d, n), where d denotes the dimension of feature and n denotes the number of training/test examples ($n = 509914$ for training while $n = 681089$ for test on the MARS dataset).

D. Ablation Studies and Model Analyses

To provide more insights on the performance of our approach, we conduct a number of ablation studies by isolating each main component, e.g., how the kNN-triplet removes the outliers in a set (video), how the P2S model outperforms the P2P model, and how the joint learning of the feature and point-to-set distance works.

How *k*NN-triplet model removes outliers. Because the iLIDS-VIDS and PRID 2011 datasets are manually selected and have no outliers, most of the algorithms based on this assumption are invalid in real-world scenarios. To fill this gap, we design four kinds of outlier to simulate real-world scenarios, i.e. occlusion, motion blur, partial body, and false alarm, as shown in Fig. 8. For each video, we randomly generate 10% outliers (denoted as PRID2011-P2S-0.1); i.e., if a video has 100 frames, we add 10 outliers (noisy frames) to the video. Each outlier is randomly selected from four kinds of outlier. In this way, we obtain three new datasets: PRID2011-P2S-0.1, PRID2011-P2S-0.2, and PRID2011-P2S-0.4. In this sub-section, we do not focus on state-of-the-art performance. Therefore, we simply pre-train on iLIDS-VIDS-P2S-0.x and fine-tune on PRID2011-P2S-0.x, but not on MARS-P2S dateset.

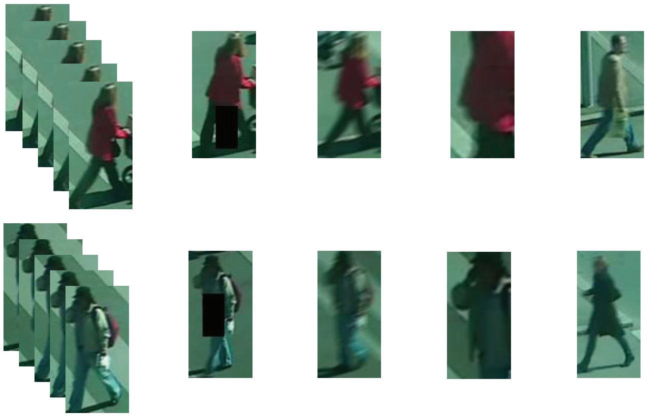
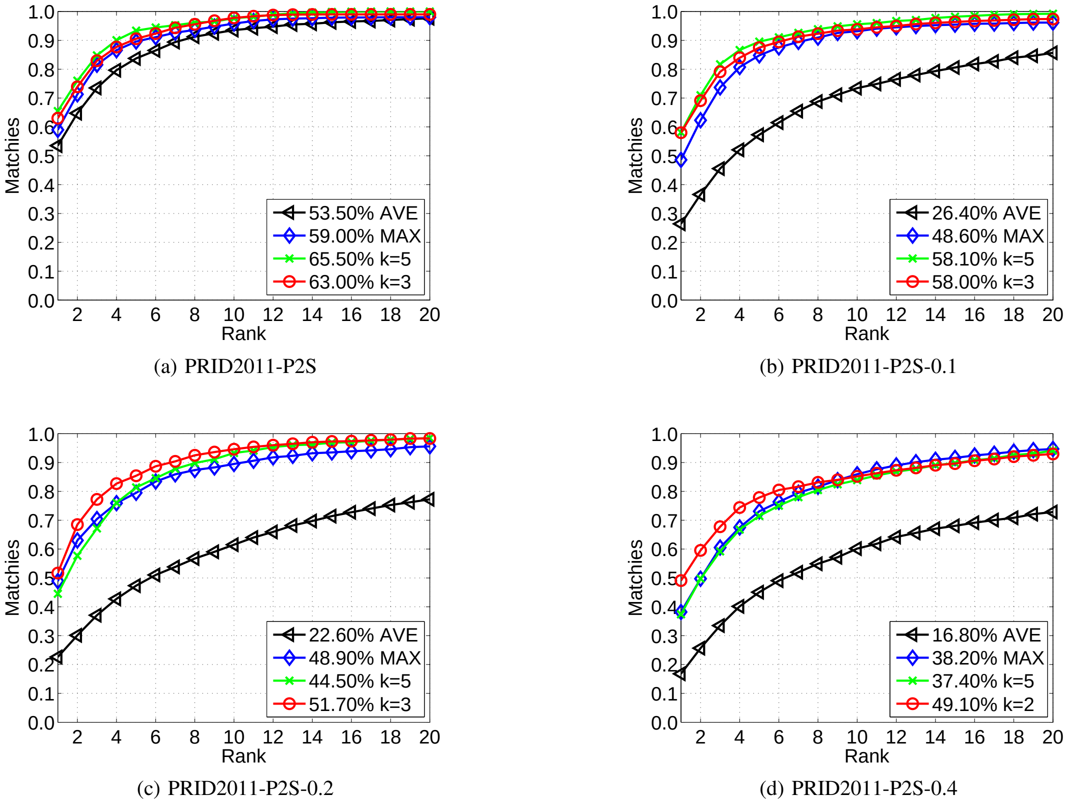


Fig. 8. Each row shows one video. Each video may contain four kinds of outlier: occlusion, motion blur, partial body and false alarm (from the second column to the fifth column).

To show how the *k*NN-triplet removes the unavailable frames for each video, we replace the *k*NN-triplet module by 1) "max temporal pooling + tripletloss"; and 2) "average temporal pooling + tripletloss". We also model the global distance metric and local distance metric in our P2SNet model by setting 3) $k = T$; and 4) $k < T$, where T is the sequence length and k is the number of nearest neighbors. In this experiment, we set $T = 5$. The results for the four datasets are reported in Fig. 9(a)-(d), respectively. We can see that without removing the outliers, the performance drops significantly. With increasing noise (Fig. 9(a)-(d)), the performance of the average temporal pooling method drops by 36.70%,that of the max temporal pooling drops by 20.80%, and the global method



(a) PRID2011-P2S

(b) PRID2011-P2S-0.1

(c) PRID2011-P2S-0.2

(d) PRID2011-P2S-0.4

Fig. 9. Results of the ablation studies demonstrating the effectiveness of removing outliers. (a) - (b) show the CMC curves of PRID2011-P2S with 0%, 10%, 20%, and 40% outliers added.

with $k = 5$ drops by 28.40%. The performance of our local method with $k = 3$ or $k = 2$ only drops by 13.90%. Note that the outliers in the PRID2011-P2S dataset have been manually removed, and thus the performance of the four methods is about the same, as shown in Fig. 9(a).

How the P2S model outperforms the P2P model. To show how the P2S model captures the available frames in the surveillance videos, we design eight experiments for the PRID2011-P2S benchmark by setting the video sequence length at 1) $T = 1$; 2) $T = 2$; 3) $T = 3$; 4) $T = 5$; 5) $T = 10$; 6) $T = 15$; 7) $T = 20$; and 8) $T = 25$. The quantitative CMC curves in the PRID2011-P2S dataset are reported in Fig. 10(a). We can see that as the video sequence length increases, the performance significantly increases from 60.50% to 67.60%, which can be explained by the richer sequence information in the longer sequence lengths. However, when the video sequence length exceeds 25 frames, the performance does not increase (as shown in Fig. 10(b)) because 25 frames randomly selected from a video contain enough information to describe a pedestrian with respect to free poses, deformation and background clutters in PRID2011-P2S.

How joint learning works. To justify the effectiveness of our joint learning method, we conduct a comparative experiment on iLIDS-VIDS-P2S, PRID2011-P2S, and MAR-P2S by 1) jointly optimizing the point-to-set metric learning and CNN feature learning using all frames, e.g., P2SNet (joint, T=ALL); and 2) using the deep feature to learn a point-to-set distance metric, e.g., “deep+XQDA”.

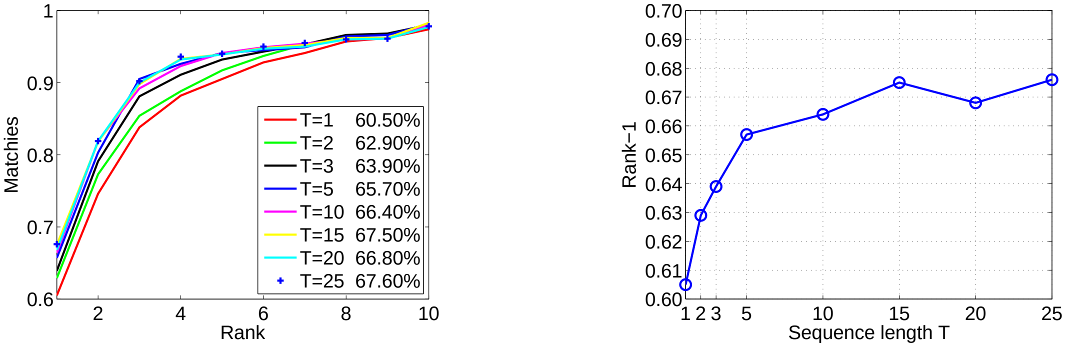
	Rank	1	5	10	20
iLIDS-VID-P2S	deep+metric	37.60	64.53	76.67	86.53
	P2SNet (Ours)	40.00	68.54	78.10	90.03
PRID2011-P2S	deep+metric	66.07	86.29	92.36	96.40
	P2SNet (Ours)	73.31	90.45	94.66	97.75
MARS-P2S	deep+metric	50.35	68.16	74.37	80.42
	P2SNet (Ours)	55.25	72.88	78.69	83.69

TABLE II
TOP MATCHING RANK(%) ON THREE DATASETS USING ALEXNET.

Our joint learning model outperforms the competing methods that use the same network to extract deep feature and then learn a point-to-set distance by 14.00%, 6.67% and 5.10% on iLIDS-VIDS-P2S, PRID2011-P2S and MARS-P2S, respectively, as shown in Fig. 7(a-c). By comparing these two methods, we can conclude that the joint learning makes sense.

E. Comparisons with Deeper Networks

MARS is a very large video-based re-id benchmark dataset. We can train a deeper network on MARS. Following [27], we conduct experiments on three datasets using AlexNet [21] for evaluation. In the experiments, we use the same convolutional neural network and implement the same experimental settings. The results are shown in Table II. The best performance are shown in red color. It is observed that our method still



(a) Performance w.r.t. T

(b) Rank-1 accuracy w.r.t. T

Fig. 10. The effectiveness of increasing the sequence length in our framework. (a) shows the CMC curves of P2SNet with different sequence lengths; (b) shows the rank-1 accuracy w.r.t. the sequence length.

outperforms the "deep+metric" models as the networks go deeper.

VI. Conclusion

In this paper, we address an image-to-video person re-id model. We integrate feature learning and point-to-set distance metric learning by building an end-to-end deep neural network architecture. The k-nearest neighbor triplet (kNN-triplet) module is used as a denoiser to remove the outliers in a video. The results of our extensive experiments demonstrate the superior performance of our model compared with other point-to-set models and related state-of-the-art methods. In future research, we plan to collect a large-scale point-to-set dataset and extend our model to more tasks.

Acknowledgments

The authors would like to thank the anonymous reviewers for providing the constructive advice on the revision of this paper.

References

[1] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in Proc. Intl Conf. Comput. Vis. IEEE, pp. 1116-1124, 2015.

[2] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan,"Object detection with discriminatively trained part-based models," IEEE Trans. Pattern Anal. and Mach. Intell., vol. 32, no. 9, pp. 1627-1645, 2010.

[3] G. Wang, L. Lin, S. Ding, Y. Li, and Q. Wang, "Dari: Distance metric and representation integration for person verification," arXiv preprint arXiv:1604.04377, 2016.

[4] L. Lin, G. Wang, W. Zuo, F. Xiangchu, and L. Zhang, "Cross-domain visual matching via generalized similarity measure and feature learning," IEEE Trans. Pattern Anal. and Mach. Intell., vol. 39, no. 6, pp. 1089-1102, 2016.

[5] S. Ding, L. Lin, G. Wang, and H. Chao, "Deep feature learning with relative distance comparison for person re-identification," Pattern Recognition, vol. 48, no. 10, pp. 2993C3003, 2015.

[6] S. Bak, S. Zaidenberg, B. Boulay, and F. Bremond, "Improving person re-identification by viewpoint cues," in Advanced Video and Signal Based Surveillance, pp. 175-180, 2014.

[7] N. Gheissari, T. B. Sebastian, and R. Hartley, "Person reidentification using spatiotemporal appearance," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit, pp. 1528-1535, 2006.

[8] K. Liu, B. Ma, W. Zhang, and R. Huang, "A spatio-temporal appearance representation for video-based pedestrian re-identification," in Proc. Intl Conf. Comput. Vis. IEEE, pp. 3810-3818, 2015.

[9] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person re-identification by discriminative selection in video ranking," , IEEE Trans. Pattern Anal. and Mach. Intell., vol. 38, no. 12, pp. 2501-2514, 2016.

[10] I. Kviatkovsky, A. Adam, and E. Rivlin, "Color invariants for person reidentification," IEEE Trans. Pattern Anal. and Mach. Intell., vol. 35, no. 7, pp. 1622-1634, 2013.

[11] J.-M. Morel, A. B. Petro, and C. Sbert, "Fast implementation of color constancy algorithms," in Proc. of SPIE, SPIE - The International Society for Optical Engineering, vol. 7241, pp. 724106-724106, 2009.

[12] A. Gijsenij, R. Lu, and T. Gevers, "Color constancy for multiple light sources," IEEE Trans. Image Process, vol. 21, no. 2, pp. 697-707, 2012.

[13] A. Mignon and F. Jurie, "Pcca: A new approach for distance learning from sparse pairwise constraints," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit, pp. 2666-2672, 2012.

[14] F. Xiong, M. Gou, O. Camps, and M. Sznaier, "Person re-identification using kernel-based metric learning methods," in Proc. Eur. Conf. Comput. Vis., Springer, pp. 1-16, 2014.

[15] W.-S. Zheng, S. Gong, and T. Xiang, "Reidentification by relative distance comparison," IEEE Trans. Pattern Anal. and Mach. Intell., vol. 35, no. 3, pp. 653-668, 2013.

[16] J. You, A. Wu, X. Li, and W.-S. Zheng, "Top-push video-based person re-identification," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit, pp. 1345-1353, 2016.

[17] R. Wang, S. Shan, X. Chen, and W. Gao, "Manifold-manifold distance with application to face recognition based on image set," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit, pp. 1-8, 2008.

[18] S. Zhou, V. Krueger, and R. Chellappa, "Probabilistic recognition of human faces from video," IEEE Computer Vision and Image Understanding, vol. 91, no. 1, pp. 214-245, 2003.

[19] Z. Huang, X. Zhao, S. Shan, R. Wang, and X. Chen, "Coupling alignments with recognition for still-to-video face recognition," in Proc. Intl Conf. Comput. Vis. IEEE, pp. 3296-3303, 2013.

[20] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," Journal of Machine Learning Research, vol. 10, pp. 207-244, 2009.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in Neural Information Processing Systems, pp. 1097-1105, 2012.

[22] L. Lin, G. Wang, R. Zhang, R. Zhang, X. Liang, and W. Zuo, "Deep structured scene parsing by learning with image descriptions," arXiv preprint arXiv:1604.02271, 2016.

[23] Y. Li, G. Wang, L. Lin, and H. Chang, "A deep joint learning approach for age invariant face verification," in Computer Vision, Springer, pp. 296-305, 2015.

[24] C. C. Aggarwal, "Proximity-based outlier detection," in Outlier Analysis, Springer, pp. 101-133, 2013.

[25] Y.-C. Chen, W.-S. Zheng, and J. Lai, "Mirror representation for modeling view-specific transform in person re-identification," in Proc. International Conference on Artificial Intelligence. AAAI Press, pp. 3402-3408, 2015.

[26] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," in Image Analysis, Springer, pp. 91-102, 2011.

[27] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian, "MARS: A Video Benchmark for Large-Scale Person Re-identification," in Proc. Eur. Conf. Comput. Vis., Springer, pp. 868-884, 2016.

[28] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance, vol. 3, no. 5. Citeseer, 2007.

[29] Z. Huang, R. Wang, S. Shan, and X. Chen, "Learning euclidean-to-riemannian metric for point-to-set classification," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit, pp. 1677-1684, 2014.

[30] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit, pp. 2197-2206, 2015.

[31] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit, pp. 2288-2295, 2012.

[32] G. Wang, P. Luo, L. Lin and X. Wang, "Learning Object Interactions and Descriptions for Semantic Image Segmentation," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit, pp. 5859-5867, 2017.

[33] S. Chen, C. Guo, J. Lai, "Deep Ranking for Person Re-Identification via Joint Representation Learning," IEEE Trans. Image Process, vol. 25, no. 5, pp. 2353-2367.

Guangcong Wang received the B.E. degree in communication engineering from Jilin University, Changchun, China, in 2015. He is currently pursuing a Ph.D. degree in the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China. His research interests are in computer vision, image processing, pattern recognition, and multiple target tracking.

Jianhuang Lai received the Ph.D. degree in mathematics in 1999 from Sun Yat-sen University, China. He joined Sun Yat-sen University in 1989 as an assistant professor, where he is currently a Professor of the School of Data and Computer Science. His current research interests are in the areas of digital image processing, pattern recognition, multimedia communication, multiple target tracking, and wavelet and its applications. He has published over 100 scientific papers in the international journals and conferences on image processing and pattern recognition, e.g., IEEE TPAMI, IEEE TNN, IEEE TKDE, IEEE TIP, IEEE TSMC (Part B), IEEE TCSVT, Pattern Recognition, ICCV, CVPR, and ICDM. He serves as a vide director of the Image and Graphics Association of China and also serves as a standing director in the Image and Graphics Association of Guangdong. He is a senior member of the IEEE.

Xiaohua Xie received a B.S. degree in mathematics and applied mathematics (2005) from Shantou University, a M.S. degree in information and computing science (2007), and a Ph.D. degree in applied mathematics (2010) from Sun Yat-sen University in China (jointly supervised by Concordia University in Canada). He is currently a Research Professor at Sun Yat-Sen University (SYSU). Prior to joining SYSU, Xiaohua was an Associate Professor at Shenzhen Institutes of Advanced Technology (SIAT), Chinese Academy of Sciences. His current research fields cover image processing, computer vision, pattern recognition, and computer graphics. He has published more than a dozen papers in the prestigious international journals and conferences. He is recognized as Overseas High-Caliber Personnel (Level B) in Shenzhen, China.