LVreID: Person Re-Identification with Long Sequence Videos

Jianing Li[1], Shiliang Zhang[1], Jingdong Wang[2], Wen Gao[1], Qi Tian[3]
[1]Peking University [2]Microsoft Research [3]University of Texas at San Antonio
{ljn-vmc, slzhang.jdl, wgao}@pku.edu.cn, jingdw@microsoft.com, qi.tian@utsa.edu

Abstract

This paper mainly establishes a large-scale Long sequence Video database for person re-IDentification (LVreID). *Different from existing datasets,* LVreID *presents many important new features.* (1) long sequences*: the average sequence length is 200 frames, which convey more abundant cues like pose and viewpoint changes that can be explored for feature learning.* (2) complex lighting, scene, and background variations*: it is captured by 15 cameras located in both indoor and outdoor scenes in 12 time slots.* (3) currently the largest size*: it contains 3,772 identities and about 3 million bounding boxes. Those unique features in* LVreID *define a more challenging and realistic person ReID task.*

Spatial Aligned Temporal Pyramid Pooling (SATPP) network is proposed as a baseline algorithm to leverage the rich visual-temporal cues in LVreID *for feature learning. SATPP jointly handles the misalignment issues in detected bounding boxes and efficiently aggregates the discriminative cues embedded in sequential video frames. Extensive experiments show feature extracted by SATPP outperforms several widely used video features. Our experiments also prove the ReID accuracy increases substantially along with longer sequence length. This demonstrates the advantage and necessity of using longer video sequences for person ReID.*

1. Introduction

Person Re-Identification (ReID) refers to the procedure of identifying a probe person in a camera network by matching his/her images or video sequences. Because of its importance in public security, person ReID has drawn lots of attention from both academia and industry. However, person ReID is challenging because different persons may exhibit similar appearances, and same person may appear differently under different cameras.

Current researches generally focus on two lines of ReID tasks that depend on single image and video, respectively. The key step of image based ReID is learning discriminative visual representations from static images. The differences between image and video make video based ReID a more complicated task. For instance, video based ReID algorithm needs to extract both visual and temporal cues and deal with issues like huge visual redundance and unequal sequence length. Video based ReID could naturally leverage the rich visual and temporal cues in surveillance videos, thus has potential to achieve better ReID performance.

Figure 1. Frames evenly sampled from sequences in *PRID* [8], *iLIDS-VID* [35], and *MARS* [45]. Each row shows two sequences of the same person. Fig. 2 shows sampled frames from *LVreID*.

Recent years have witnessed impressive progresses in image based person ReID, *e.g.*, deep visual representations have significantly boosted the ReID performance on image datasets [16, 46]. The advantages of fusing visual and temporal cues for video based person ReID has not been fully investigated and demonstrated. The possible reasons could be due to several limitations existing in current person ReID video datasets. First, the lengths of video sequences in existing datasets are too short to provide enough temporal information than static images. For instance, the average length of video sequences is 58 frames in *MARS* [45] and 73 frames in *iLIDS-VID* [35], respectively. This means those sequences only last for 2 to 3 seconds. Fig. 1 illustrates several video sequences in *MARS* [45], *iLIDS-VID* [35], and *PRID* [8]. It is obvious that, the person image in the end of each sequence shows similar appearances with the beginning image, thus cannot provide extra complementary cues about the person identity. Moreover, this issue may hinder the research efforts on video feature learning for person ReID because the video sequence may convey similar cues with a static key frame.

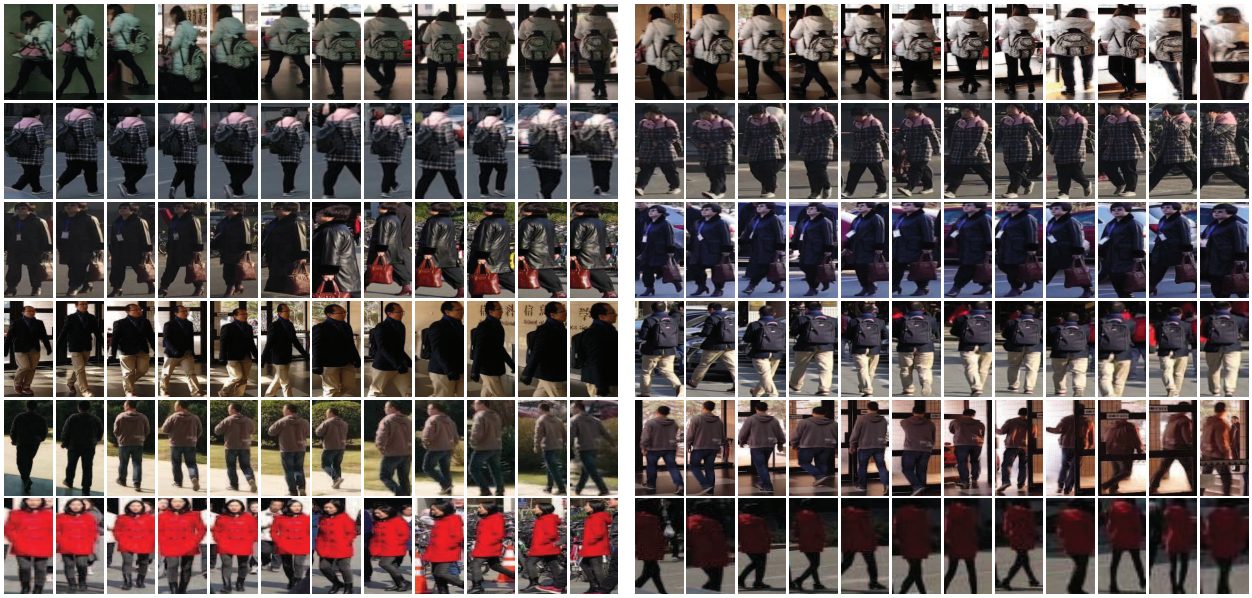

Figure 2. Frames evenly sampled from sequences in *LVreID*. Each row shows two sequences of the same person. Compared with existing datasets, *LVreID* is more challenging and provides more abundant visual and temporal cues.

Another issue of existing datasets is the limited data size. The currently largest video ReID dataset, *i.e.*, *MARS* [45], contains less than 1500 labeled identities. Also, person ReID tasks in real scenarios are more challenging than the ones defined in current datasets. Videos in real scenarios could be collected for multiple days by a camera network deployed in different scenes. In contrast, most of existing video datasets are collected in constrained environments with limited number of cameras, single time slot, fixed scene and lighting condition. With the ability of exploring more extra cues, video based ReID has potential to perform better than image based ReID in those challenging tasks. The insufficiency of video sequence length and identity scale in current datasets make it hard to investigate the advantages of video based person ReID.

Dataset: This paper is motivated to overcome those issues in current datasets and facilitate the research towards more realistic person ReID task. The collected Long sequence Video person re-IDentification (*LVreID*) dataset presents several important features. 1) The average sequence length in *LVreID* is 200 frames, which is significantly longer than the ones in previous datasets. Fig. 2 illustrates several sampled frames in video sequences on *LVreID*. It can be observed that, within the long sequences, there commonly exist substantial pose changes, viewpoint variations, and lighting changes, *etc*. Those variations inside sequences imply richer visual and temporal cues, which could be helpful for person ReID. 2) *LVreID* is currently the largest and most realistic video ReID dataset. It is constructed from 180 hours of videos taken by both indoor and outdoor cameras during multiple time slots over a month. It also annotates 3,772 identities and nearly 3 million bounding boxes. 3) *LVreID* generates high quality video sequences by using Faster RCNN [25] for bounding box detection and robust deep features for bounding boxes matching among adjacent video frames. Therefore, *LVreID* defines a challenging and realistic video based person ReID task. It is also more reliable than previous person ReID video databases, because it allows algorithms to explore more abundant cues conveyed in long video sequences.

Baseline Solution: Based on *LVreID*, we further study the learning of discriminative video representation for person ReID. Existing algorithms commonly generate video features by average pooling frame features [45], or applying the Long Short-Term Memory (LSTM) network [39, 9] to capture temporal cues. Average pooling treats each frame equally and may lose important temporal cues. LSTM model is complicated for training, especially for the long sequence. Besides that, spatial misalignment commonly exists in detected bounding boxes. In video sequences, spatial misalignment may cause sudden foreground variations between adjacent frames and degrade the robustness of learned sequence features.

To address the above issues, we propose a Spatial Aligned Temporal Pyramid Pooling (SATPP) network as a baseline solution for video feature learning in *LVreID*. SATPP first aligns person images by imposing an 2D affine transformation on each video frame. Deep features extracted on the aligned frames are then processed by a temporal pyramid pooling layer to fuse both the long and short term deep features. Therefore, SATPP jointly aligns person images and learns video features from sequences with

unequal length.

Extensive experiments are carried out on *LVreID* using different features. We observe that, the performances of video features are substantially boosted with long video sequences. This clearly demonstrates the advantages of using long sequences in person ReID. We also compare the SATPP with other video feature learning strategies on *LVreID* and three other commonly used video ReID datasets. Experiments show that SATPP presents higher accuracy and lower complexity than existing strategies like LSTM [39, 9].

Contributions: The contributions of this work can be summarized into three folds: 1) a more challenging large-scale *LVreID* dataset is annotated and will be released, 2) an efficient SATPP network is proposed for video feature learning, and 3) the proposed dataset and network have potential to facilitate the future research on discriminative video feature learning for video based person ReID.

2. Related work

Existing person ReID works can be summarized into two categories, *i.e.*, image based person ReID and video based person ReID, respectively. This section briefly reviews those two categories of works.

Lots of image based person ReID works have been published in recent years. Early works basically carry on two important steps: a) learning discriminative image representations [19, 44, 35, 28] and b) learning discriminative distance metrics for image feature matching [22, 24, 38, 20]. The release of large-scale ReID datasets like *CUHK03* [16] and *Market-1501* [46] makes training deep models for person ReID feasible.

Many researchers have leveraged deep models in person ReID by learning deep feature representations [37, 45, 1, 29, 30, 28, 41] and distance metrics [4, 33, 6, 40]. Existing works usually extract deep feature representations for person images from convolutional layers [1, 16, 37, 6] or Fully Connected (FC) layers [42, 27, 45]. Some works first learn deep representations with the Triplet Loss or the Siamese Loss, then utilize Euclidean distance for feature comparison [34, 2, 4, 33].

Many research efforts have been conducted on video based person ReID. Some works extract space-temporal cues such as 3D-SIFT [26] and HOG3D [12] to build the video representations. Those hand-crafted representations present limited robustness and discriminative power when compared with deep features [45]. Recently, many works first extract deep features from video frames, then accumulate frame features as video features [39, 36, 21, 23, 45]. Some works apply average pooling for video feature generation [45]. Some others apply the Recurrent Neural Networks (RNN) to accumulate frame features into video features [36, 21, 23]. For example, Yan *et al*. [39] utilize the LSTM [9] network to learn the temporal cues in videos sequences.

From the above reviews, it can be observed that, most of research efforts focus on image based person ReID. Therefore, more in-depth research should be conducted to demonstrate the advantages of video based person ReID. The *LVreID* is proposed to overcome the limitations in current datasets. The SATPP considers the misalignment issue in video feature extraction and fuses the image-level features in a more reasonable and efficient way. Those contributions make this work novel from previous ones.

3. The *LVreID* dataset

3.1. Overview of Existing Datasets

We summarize existing datasets for video based person ReID in Table 1. As shown in the table, *PRID*-2011 [8] and *iLIDS-VID* [35] contain 200 and 300 identities, respectively, and each identity has 2 video sequences captured by two different cameras. *MARS* [45] is currently the largest video dataset and contains 1,261 identities, 20,715 sequences recorded by 6 outdoor cameras. The video sequences in *PRID* and *iLIDS-VID* are manually generated with hand drawn bounding boxes. The bounding boxes on *MARS* are generated with DPM detector [5]. According to the Table 1 we can briefly summarize the limitations in existing video based person ReID datasets: 1) short sequence length, 2) limited scale and variations compared with data in real scenarios, and 3) the sequences are generated either with expensive hand annotation or outdated detectors. Those limitations make it necessary to collect a larger and more realistic video dataset for person ReID.

3.2. Description for LVreID

Video Capture: The collection procedure of *LVreID* is carefully designed to simulate the real scenario as much as possible. We utilize a camera network containing 12 outdoor cameras and 3 indoor cameras for data collection. In this camera network, 13 cameras record 1080×1920 HD videos with 30 Frames Per Second (FPS). The other 2 cameras record 1080×1920 HD videos with 50 FPS. 4 days during January to March in 2017 are selected for data recording. For each day, 3 hours of videos taken in the morning, noon, and afternoon, respectively, are selected for pedestrian detection and annotation. Our raw data thus contains 180 hours of HD videos, 12 outdoor cameras, 3 indoor cameras, and 12 time slots with different lighting conditions.

Sequence Detection and Annotation: Faster RCNN [25] is utilized for pedestrian bounding box detection on each video frame. After bounding box detection, we hence design a sequence extraction strategy to generate sequences as long as possible. For each camera, we first detect the appearance of a pedestrian. This pedes-

Table 1. Comparison between *LVreID* and existing video based person ReID datasets.

dataset	identities	sequences	bboxes	# of frames	cams indoor	cams outdoor	evaluation
LVreID	**3,772**	**14,943**	**2,989,436**	**200**	**3**	**12**	**CMC + mAP**
MARS [45]	1,261	20,715	1,067,516	58	0	6	CMC + mAP
PRID-2011 [8]	200	400	40,033	100	0	2	CMC
iLIDS-VID [35]	300	600	42,460	73	2	0	CMC

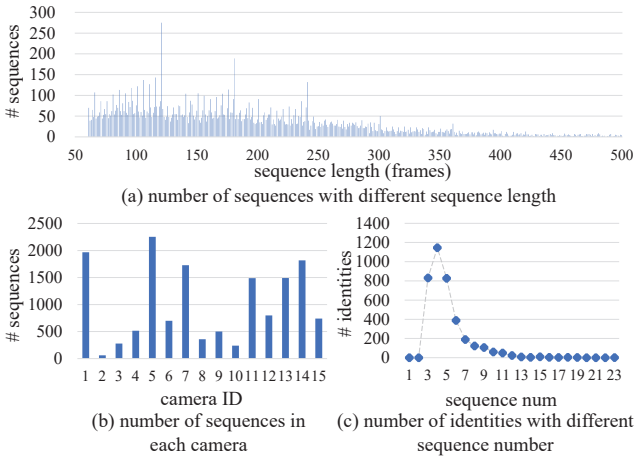

Figure 3. Some statistics on *LVreID* dataset.

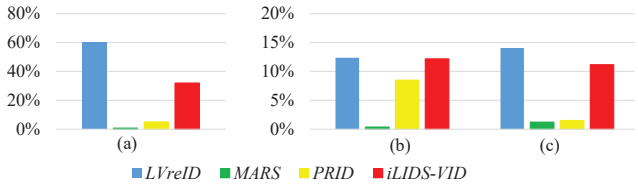

Figure 4. (a) shows the portion of identities whose sequences present substantially lighting changes. (b) and (c) show the portion of sequences, within which frames present substantial viewpoint changes (>60°) and background changes, respectively. This statistic randomly samples 300 identities in *LVreID* and *MARS*, and uses all identities in *PRID* and *iLIDS-VID*.

trian is then tracked by matching his/her bounding boxes among adjacent video frames with deep features [28]. The tracking is finished when this pedestrian leaves this camera or the matching similarity drops below a threshold. After discarding some sequences with too short length, we finally collect 14,943 sequences of 3,772 pedestrians, and the average sequence length is 200 frames. The person identity annotation is finished manually by three labelers for two months.

Statistics and Comparison: Some statistics on *LVreID* are shown in Fig. 3. Fig. 4 and Table 1 compare *LVreID* with existing video based person ReID datasets. In this comparison, we conclude that *LVreID* shows the following important features:

1) *Longer sequences.* We can see from Fig. 3(a) that, most of sequences in *LVreID* contain 100 to 250 frames. As shown in Fig. 2, long sequence presents abundant visual and temporal cues like pose and viewpoint changes, and would benefit future research on feature learning in person ReID.

2) *More accurate pedestrain tracklets.* In each camera, pedestrians are tracked as long as possible by Faster RCNN [25] detector and bounding box matching with deep features. This strategy gets accurate long sequences and is easy to repeat in real systems.

3) *Currently the largest video dataset for person ReID.* *LVreID* contains significantly larger number of identities and bounding boxes. For example, the numbers of identities and bounding boxes in *LVreID* are three times of that in previously largest video dataset, *i.e.*, the *MARS*. This encourages the research on more efficient person ReID algorithms.

4) *More realistic and challenging person ReID task.* As shown in Fig. 4, *LVreID* is carefully collected to guarantee variations of lightings, backgrounds, scenes, viewpoints, *etc*. As shown in Fig. 3, most identities have 4-5 sequences with different appearance cues. Those challenges shift the research efforts towards the real application of person ReID.

Therefore, we conclude that *LVreID* defines a more reliable video based person ReID task, thus has potential to move forward the research on video based person ReID.

3.3. Evaluation Protocol

LVreID contains about 3 millions of bounding boxes and 3,772 identities. It is time consuming to randomly select training and testing subsets. Therefore, we provide the training set and testing set. We evenly divide the 3,772 pedestrians into training and testing sets, making both of those two sets contain 1,886 pedestrians. In the testing set, we randomly select 2364 sequences as queries and remaining 7371 sequences as gallery.

Similar to other datasets [8, 35, 45], *LVreID* treats person ReID as a cross-camera video retrieval problem. The widely used Cumulated Matching Characteristics (CMC) curve is used as evaluation metric. For each query sequence, multiple true positives could be returned. Only using CMC curve is not accurate enough to reflect the ReID accuracy. Therefore, we also use mean Average Precision (mAP) as the evaluation metric.

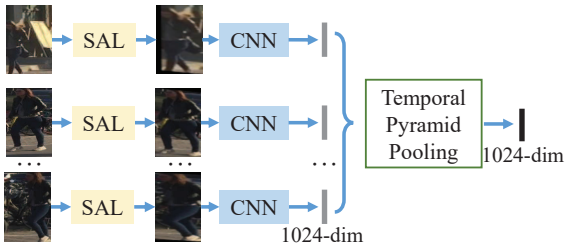

Figure 5. Illustration of the Spatial Aligned Temporal Pyramid Pooling (SATPP) model. SATPP takes a series of frames as input and outputs one video feature with fixed dimensionality.

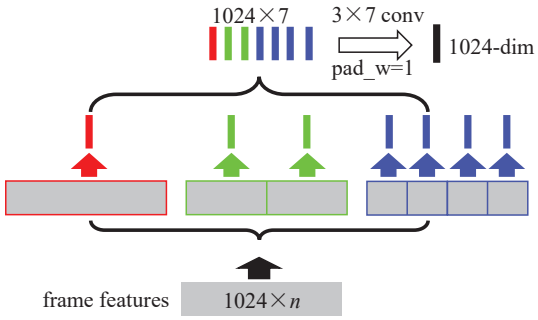

Figure 6. Illustration of Temporal Pyramid Pooling (TPP) with L=3 and d=1024.

4. Spatial Aligned Temporal Pyramid Pooling Network

LVreID raises an important problem, *i.e.*, how to extract discriminative video features by exploiting rich visual and temporal cues in long video sequences. This work proposes the Spatial Aligned Temporal Pyramid Pooling (SATPP) network as an efficient and easy to repeat video feature learning baseline. Fig. 5 shows the framework of SATPP. SATPP is proposed with two motivations: 1) handle the misalignment in detected pedestrian bounding boxes and 2) extract and fuse discriminative cues conveyed by sequential video frames.

As shown in Fig. 5, SATPP takes a video sequence as input. Each video frame is first aligned with a 2D affine transformation learned in a Spatial Alignment Layer (SAL). Then each aligned frame is fed into a CNN for frame feature extraction. A Temporal Pyramid Pooling (TPP) layer finally fuses multiple frame features into a fixed length video feature. The following parts present more details of SAL, TPP. Details of SATPP training are given in Sec. 5.3.

4.1. Spatial Alignment Layer

The misalignment can be corrected by training another pedestrian detector on outputs of Faster RCNN. However, this strategy makes the network complicated and expensive to compute. Inspired by the Spatial Transformer Networks (STN) [10], we propose to align an image by predicting an affine transformation.

We choose 2D affine transformation and learn an 6-dimensional parameter A_θ for the alignment, *i.e.*,

$$\begin{pmatrix} x^s \\ y^s \end{pmatrix} = A_\theta \begin{pmatrix} x^t \\ y^t \\ 1 \end{pmatrix} = \begin{bmatrix} \theta_1 & \theta_2 & \theta_3 \\ \theta_4 & \theta_5 & \theta_6 \end{bmatrix} \begin{pmatrix} x^t \\ y^t \\ 1 \end{pmatrix}, \quad (1)$$

where the θ is an affine parameter, (x^t, y^t) are target coordinates of output image and (x^s, y^s) are source coordinates of input image. With learned affine parameters, person images will be shifted, rotated, and resized to generate a better aligned image.

The STN consists of three components: localization network, parameterized sampling grid, and differentiable image sampling. We follow the structure of STN to learn affine transformation parameters. More details of STN can be found in [10]. As shown in our experiments, aligned person images helps to learn more robust and discriminative video features.

4.2. Temporal Pyramid Pooling Layer

The aligned frames are fed into CNN for deep feature extraction. Details of the utilized CNN will be presented in Sec. 5.3. Because video sequences may have different lengths, we proceed to convert the frame-level features into a video-level feature with fixed dimensionality. Traditional works usually use two types of pooling strategies, *i.e.*, average-pooling and max-pooling, which compute the averaged value and max value on each feature dimension, respectively. Note that, different pooling strategies are suited for different types of features, *e.g.*, max-pooling is generally suited for sparse features. It is difficult to decide which pooling strategy is optimal for our task. Moreover, most of pooling strategies discard the temporal clues, which might be important to distinguish person identities.

We thus propose the TPP Layer to seek a more reasonable pooling solution. TPP is inspired by the Pyramid matching [14] and the recent Spatial Pyramid Pooling [7], which pool visual features in spatial grids with different scales to preserve discriminative spatial information. Suppose there are n frame features in a sequence with dimension d, TPP builds a L-layer temporal pyramid by evenly dividing the n frames into 2^{i-1} segments in the i-th layer. This procedure is illustrated in Fig. 6. Average pooling each segment on the L layers results in a feature F with fixed dimension D, *i.e.*,

$$D = d \times (2^L - 1). \quad (2)$$

Because the dimensionality of F could be high with large L, we further learn a more compact descriptor on it. To make the TPP layer efficient for training and testing, we

avoid introducing to many parameters into it. We resize F into an 2-D feature map with width 2^L-1 and height d. An d dimensional feature thus could be generated by learning a $w \times 2^L - 1, (w << d)$ sized convolutional kernel, *i.e.*,

$$\bar{F} = F \otimes W, F \in \mathbb{R}^{d \times 2^L - 1}, W \in \mathbb{R}^{w \times 2^L - 1}, \quad (3)$$

where $\bar{F}$ denotes the final video feature and W denotes the convolutional kernel need to be learned. We have experimented different settings of w and find w=3 gets the best performance. We thus fix w as 3.

Fig. 6 illustrates the TPP layer with L=3 and d=1024, where the final feature is generated from an 1024×7 sized feature map. It is easy to infer that, TPP only involves $3 \times (2^L-1)$ parameters, thus is efficient for training and testing. The validity of TPP will be shown in our experiments.

5. Experiment

5.1. Dataset

Besides of *LVreID*, we select three widely used datasets as our evaluation groundtruths, including *PRID*-2011 [8], *iLIDS-VID* [35] and *MARS* [45].

PRID-2011 dataset consists of 400 sequences of 200 pedestrians from two cameras. Each sequence has a length between 5 and 675 frames. Following the implementation in previous works [39, 35], we choose sequences containing more than 21 frames, and evenly divide them into training and testing sets.

iLIDS-VID consists of 600 sequences of 300 pedestrians from two non-overlapping cameras. Each sequence has a variable length between 23 and 192 frames. We also evenly divide this dataset to use 150 pedestrians for training and 150 pedestrian for testing, respectively.

MARS consists of 1261 pedestrians and 20,715 sequences under 6 cameras. Each pedestrian is captured by at least 2 cameras. This dataset provides fixed training and testing sets, which contain 630 and 631 pedestrians, respectively.

5.2. Compared Features

LOMO feature: Local Maximal Occurrence Feature (LOMO) [18] is a competitive hand-crafted feature which is shown robust against the variations in person ReID. All images are resized to a fixed size to extract LOMO feature. Then a pooling operation is imposed to get a sequence level representation.

CNN feature: Convolutional Neural Network (CNN) has demonstrated promising performance in image based person ReID [37, 45, 1, 29, 30, 28, 41]. We follow previous works [1, 37, 45, 6, 28] and train a deep CNN model with a classification task. We hence extract an 1024-dim feature with this CNN model as image-level representation. Different pooling operations will be applied on CNN feature to generate the sequence level features.

LSTM feature: Many previous works [36, 21, 23, 9, 39] use the recurrent model for video based person ReID. Those works show features produced by recurrent models perform significantly better than hand-crafted features. We also implement a standard LSTM network to capture the temporal information.

5.3. Implementation Details

For the CNN training, we build our network based on the GoogLeNet [31] and use parameters pre-trained on Imagenet [3] to initialize our CNN network. We remove the three loss branches in GoogLeNet and impose an 1×1 conv layer and a global pooling layer for classification. All input images are resized to 128×64, the mean value is subtracted from each color channel. And each batch contains 128 images. The initial learning rate is set as 0.01, and is gradually lowered after 1×10^4 iterations. It should be noted that, the learning rate in SAL network is only 0.1% of that in feature learning network. During testing, we employ a Global Average Pooling (GAP) layer after the "inception_5b" layer [31] to extract an 1024-dim feature.

We use the 1024-dim CNN feature as input to train the TPP layer. 64 frames sampled from an original sequence compose a training sequence, and each training batch contains 8 sequences. The initial learning rate is set as 0.001 and decreases after 3×10^4 iterations. We train the TPP with totally 5×10^4 iterations.

To make a comparison, we train a CNN+LSTM network on *LVreID* following similar structures in [36, 23, 21]. We directly use the LSTM model in Caffe [11]. Each training batch contains 8 sequences, where each consists of 32 sampled frames. The output of each LSTM block is an 1024-dim feature. The initial learning rate is set as 0.001 and decreases after 3×10^4 iterations. The training process totally contains 5×10^4 iterations.

All of our deep models are trained and fine-tuned on Caffe [11] with GTX TITAN X GPU, Intel i7 CPU, and 128GB memory.

5.4. Evaluation of SATPP

This section evaluates the validity of SAL and TPP in SATPP. We compare features learned with SATPP against three other features, *i.e.*, LOMO, CNN feature, and CNN feature extracted after SAL, respectively. We present the comparisons in Table 2.

Table 2 shows the performance of LOMO feature on 3 datasets. We can see that, on the three datasets, LOMO feature gets the best performance on *PRID*, which is small and presents relatively stable lighting conditions and backgrounds. However, on the large dataset *MARS* and *iLIDS-VID* which contains substantial variations, the performance

Table 2. The performance of different features and pooling strategies on the *PRID*, *iLIDS-VID* and MARS. "image" denotes the static image feature extracted from the first image in a sequence. TPP with L=3 is also tested in this experiment.

dataset		*PRID*				*iLIDS-VID*				*MARS*				
feature	fuse method	r1	r5	r10	r20	r1	r5	r10	r20	mAP	r1	r5	r10	r20
LOMO	image	29.21	58.43	75.28	84.27	8.00	22.00	30.00	39.33	6.07	17.68	28.38	34.70	41.52
	max-pool	24.72	47.19	68.54	86.52	4.00	12.00	19.33	24.00	6.26	17.83	30.35	43.23	43.38
	avg-pool	47.19	70.79	77.53	86.52	12.67	22.67	29.33	40.67	9.54	23.74	37.12	43.23	48.69
CNN	image	62.92	88.67	96.63	98.88	26.67	52.67	63.33	74.67	35.00	52.58	72.53	80.15	84.70
	max-pool	75.28	97.75	98.88	100	52.00	76.67	86.67	92.00	44.74	65.91	81.41	86.21	90.05
	avg-pool	77.53	97.75	100	100	53.33	77.33	88.67	93.33	51.47	67.08	84.65	89.34	92.12
	TPP	78.65	98.88	100	100	54.67	78.67	88.67	93.33	51.95	68.54	84.70	89.24	91.77
CNN+SAL	image	64.04	88.67	97.75	98.88	28.00	60.00	72.00	84.67	35.64	53.43	73.23	79.75	85.96
	max-pool	79.77	98.88	100	100	54.67	78.00	89.33	95.33	50.06	68.69	84.09	88.38	92.29
	avg-pool	80.90	100	100	100	55.33	78.67	88.67	96.67	52.00	68.79	84.55	88.73	92.29
SATPP	TPP	**82.02**	**100**	**100**	**100**	**56.67**	**78.67**	**90.00**	**96.67**	**52.55**	**69.69**	**84.65**	**89.34**	**92.77**

Table 3. Comparisons with recent works on *PRID* and *iLIDS-VID*.

Dataset	*PRID*			*iLIDS-VID*		
Method	r1	r5	r20	r1	r5	r20
DFCP [17]	51.60	83.10	95.50	34.30	63.30	84.40
RFA-Net [39]	58.20	85.80	97.90	49.30	76.80	90.00
STFV3D [13]	64.10	87.30	92.00	44.30	71.70	91.70
DRCN [36]	69.00	88.40	96.40	46.10	76.80	89.70
RCN [23]	70.00	90.00	97.00	**58.00**	**84.00**	96.00
IDE [45]	77.30	93.50	99.30	53.00	81.40	95.10
SATPP	**82.02**	**100.00**	**100.00**	56.67	78.67	**96.67**

Table 4. Comparison with recent works on *MARS*.

Method	mAP	r1	r5	r10	r20
BoW+kissme [45]	15.50	30.60	46.20	-	59.20
IDE [45]	42.40	60.00	77.90	-	87.90
IDE+kissme [45]	45.60	65.00	81.10	-	88.90
LCAR [43]	-	55.50	70.20	-	80.20
CDS [32]	-	68.20	-	-	-
SFT [47]	50.70	70.60	**90.00**	-	**97.60**
DCF [15]	**56.05**	**71.77**	86.57	-	93.08
SATPP	52.55	69.69	84.65	89.34	92.77

of LOMO drops considerably. We thus conclude that the hand-crafted features are not discriminative and robust enough for person ReID.

Table 2 also shows that, CNN feature outperforms the LOMO feature by large margins. With average pooling, CNN feature gets 77.53% rank1 accuracy on *PRID*, which is nearly 2 times higher than that of LOMO. It also can be observed that, sequence level feature consistently outperforms the static image feature. On *iLIDS-VID*, CNN feature with average pooling achieves 53.33% rank1 accuracy, which significantly outperforms the CNN static image feature by 26.66%.

Sequence level feature is also generated with TPP. It is clear that, TPP constantly outperforms average pooling and max pooling on three datasets, *e.g.*, TPP outperforms average pooling by 1.5% in mAP on *MARS*. CNN+SAL denotes the CNN features extracted after SAL. It is clear that, SAL consistently improves the ReID performance. With average pooling, CNN+SAL outperforms CNN by 7.37%, 2.0%, and 1.71% in rank1 accuracy on *PRID*, *iLIDS-VID*, and *MARS*, respectively. This shows that the well-aligned bounding boxes promote the ReID performance.

"SATPP" in Table 2 denotes the video feature generated by our algorithm. It substantially outperforms all other features. Therefore, we conclude that our SATPP is more effective in learning video features.

5.5. Comparison with Recent Work

In Table 2, feature learned with SATPP substantially outperforms the other features. We proceed to compare SATPP with some state-of-the-art approaches. The comparisons on *PRID* and *iLIDS-VID* are shown in Table 3. From the results, we can see that SATPP gets promising performance on those two datasets. On *PRID*, SATPP performs better than all compared methods. It achieves rank1 accuracy of 82.02%, outperforming recent works DFCP [17] and RFA-Net [39] by more than 20% on rank1 accuracy. SATPP also performs better than most of the compared works on *iLIDS-VID*.

The comparison on *MARS* dataset is shown in Table 4. SATPP also presents competitive performance on rank1 accuracy and mAP. IDE [45] directly generates video feature by average pooling CNN features. SATPP performs substantially better than IDE [45] by 9.69% on rank1 accuracy. SATPP also outperforms several works published in 2017 [43, 32] and performs better than the recent SFT [47] in mAP. DCF [15] performs better than SATPP. However, DCF [15] utilizes extra human body part cues for feature extraction. Note that, SATPP modifies the GoogLeNet by inserting SAL and TPP, which only involve 6 and 21 new parameters, respectively. Therefore, we conclude that SATPP presents promising performance with an efficient network structure, thus could be a reliable baseline for *LVreID*.

Table 5. Parameter Analysis of L using SATTP on *LVreID*

L	mAP	r1	r5	r10	r20
2	49.03	63.36	81.18	**86.59**	90.39
3	**49.24**	**63.66**	**81.35**	86.29	**90.52**
4	48.98	63.32	81.10	86.25	89.93
avg-pool	48.12	62.35	79.82	85.53	90.06

Table 6. The results on the *LVreID* dataset.

method	fuse method	mAP	r1	r5	r10	r20
LOMO	image	1.50	2.41	4.99	7.15	10.19
	max-pool	0.77	1.02	3.09	4.74	7.89
	avg-pool	2.10	2.96	7.87	10.62	15.06
CNN	image	22.93	31.90	53.76	63.16	71.45
	max-pool	39.99	53.68	74.62	81.43	86.38
	avg-pool	47.13	61.38	79.21	85.01	89.93
	TPP	48.03	62.48	80.16	85.53	89.93
LSTM	max-pool	41.75	52.96	75.72	80.80	89.89
	avg-pool	45.47	56.85	78.64	84.65	90.06
CNN+SAL	image	23.60	32.06	54.65	64.64	72.47
	max-pool	41.94	54.91	77.07	83.25	89.17
	avg-pool	48.12	62.35	79.82	85.53	90.06
SATPP	TPP	**49.24**	**63.66**	**81.35**	**86.29**	**90.52**

5.6. Evaluation with the *LVreID* dataset

This section further verifies the performance of SATPP on *LVreID*. We first check the performance of SATPP with different L, which controls the number of pyramid layers in TPP. Table 5 summarizes the performance of SATPP with different L. As we can see that, $L = 3$ generally gets the best performance on the *LVreID*. The performance starts to drop if the value of L is too large. Therefore, we fix L=3 in our experiments.

The comparison between SATPP and other features on *LVreID* is shown in Table 6. We can observe that, LOMO feature gets the worst performance, which is consistent with the observation in Table 2. The CNN feature is more robust than LOMO, thus gets significantly better performance than LOMO. LSTM is commonly used for video feature extraction for person ReID. With max pooling, LSTM feature outperforms CNN feature.

It is also obvious to observe that, TPP performs better than average pooling and max pooling on *LVreID*. CNN+SAL performs image alignment before feature extraction. It also consistently outperforms CNN features. This observation shows that image alignment is necessary for bounding boxes detected by Faster RCNN. It is clear that, SATPP feature outperforms all the other features in Table 5. This validates the advantage of SATPP in video feature learning. Compared with its performance on *LVreID*, SATPP performs significantly better on *PRID*, *iLIDS-VID*, and *MARS*. We thus conclude that, *LVreID* is more challenging than existing datasets.

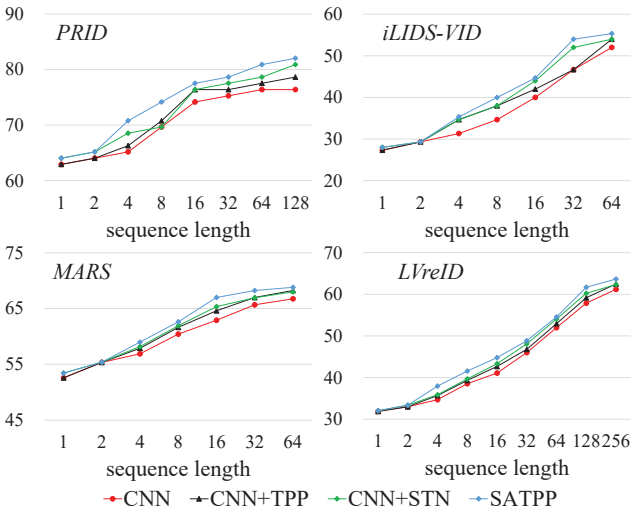

Figure 7. The rank1 accuracies of four features extracted from sequentially sampled frames with different length.

5.7. The Advantage of Long Sequence

The experimental results in Table 2 and Table 6 show that sequence level feature performs better than static image feature. We thus further evaluate the performance of features extracted with different sequence length.

From the beginning of each sequence, we sequentially sample different numbers of frames for feature extraction. The experimental results are illustrated in Fig. 7. The average sequence length on *PRID*, *iLIDS-VID*, *MARS*, and *LVreID* is 100, 73, 58 and 200, respectively. We thus set the maximum number of sampled frames as 128, 64, 64 and 256 for those datasets, respectively. We directly use the whole sequence, if its length is shorter than the sample length.

Fig. 7 compares the performance of four features. Note that, sequence length = 1 equals to image based person ReID. It is clear that, the performance of different features is substantially boosted with longer sequences. This clearly indicates the advantage of video based person ReID, *i.e.*, longer sequence provides more abundant cues that could be helpful for person ReID. Another phenomenon we observe is that, with longer sequences the performance improvement on *LVreID* is more substantial than that of the other datasets. This shows the advantage of our dataset, *i.e.*, contains longer video sequences and may facilitate future research on video feature learning for person ReID.

6. Conclusion

This paper mainly presents *LVreID*, a large-scale long sequence video database for person ReID. *LVreID* is collected to present many new features: 1) it contains longer video sequences, 2) it presents more accurate pedestrian tracklets, 3) it is currently the largest video dataset for per-

son ReID, and 4) it defines a more realistic and challenging person ReID task. Our experiments prove that the ReID accuracy increases along with longer sequence length. This validates the advantage of video based person ReID, and also demonstrates the necessity of using long sequences for person ReID in *LVreID*.

We also propose SATPP as baseline method to leverage the rich visual-temporal cues in *LVreID* for feature learning. SATPP jointly handles the misalignment issues in detected bounding boxes and efficiently aggregates the discriminative cues embedded in sequential video frames. Extensive experiments show features extracted by SATPP substantially outperform several commonly used video features. Thus, SATPP can be an efficient and effective baseline for *LVreID*.

References

[1] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *CVPR*, 2015. 3, 6

[2] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *CVPR*, 2016. 3

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009. 6

[4] S. Ding, L. Lin, G. Wang, and H. Chao. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 48(10):2993–3003, 2015. 3

[5] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2010. 3

[6] M. Geng, Y. Wang, T. Xiang, and Y. Tian. Deep transfer learning for person re-identification. *arXiv preprint arXiv:1611.05244*, 2016. 3, 6

[7] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1904–1916, 2015. 5

[8] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In *Image Analysis*, pages 91–102. Springer, 2011. 1, 3, 4, 6

[9] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 2, 3, 6

[10] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *NIPS*, 2015. 5

[11] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM MM*, 2014. 6

[12] A. Klaser, M. Marszałek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *BMVC 2008-19th British Machine Vision Conference*, pages 275–1. British Machine Vision Association, 2008. 3

[13] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2288–2295. IEEE, 2012. 7

[14] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer vision and pattern recognition, 2006 IEEE computer society conference on*, volume 2, pages 2169–2178. IEEE, 2006. 5

[15] D. Li, X. Chen, Z. Zhang, and K. Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 384–393, 2017. 7

[16] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014. 1, 3

[17] Y. Li, L. Zhuo, J. Li, J. Zhang, X. Liang, and Q. Tian. Video-based person re-identification by deep feature guided pooling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 39–46, 2017. 7

[18] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2015. 6

[19] C. Liu, S. Gong, C. C. Loy, and X. Lin. Person re-identification: what features are important? In *ECCV*, 2012. 3

[20] C. Liu, C. C. Loy, S. Gong, and G. Wang. Pop: Person re-identification post-rank optimisation. In *ICCV*, 2013. 3

[21] H. Liu, Z. Jie, K. Jayashree, M. Qi, J. Jiang, S. Yan, and J. Feng. Video-based person re-identification with accumulative motion context. *arXiv preprint arXiv:1701.00193*, 2017. 3, 6

[22] A. J. Ma, P. C. Yuen, and J. Li. Domain transfer support vector ranking for person re-identification without target camera label information. In *ICCV*, 2013. 3

[23] N. McLaughlin, J. Martinez del Rincon, and P. Miller. Recurrent convolutional network for video-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1325–1334, 2016. 3, 6, 7

[24] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian. Local fisher discriminant analysis for pedestrian re-identification. In *CVPR*, 2013. 3

[25] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 2017. 2, 3, 4

[26] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional sift descriptor and its application to action recognition. In *Proceedings of the 15th ACM international conference on Multimedia*, pages 357–360. ACM, 2007. 3

[27] H. Shi, Y. Yang, X. Zhu, S. Liao, Z. Lei, W. Zheng, and S. Z. Li. Embedding deep metric for person re-identification: A study against large variations. In *ECCV*, 2016. 3

[28] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian. Pose-driven deep convolutional model for person re-identification. *ICCV*, 2017. 3, 4, 6

[29] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian. Deep attributes driven multi-camera person re-identification. In *ECCV*, 2016. 3, 6

[30] C. Su, S. Zhang, J. Xing, Q. Tian, and W. Gao. Multi-type attributes driven multi-camera person re-identification. *Pattern Recognition*, 2017. 3, 6

[31] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015. 6

[32] Y. T. Tesfaye, E. Zemene, A. Prati, M. Pelillo, and M. Shah. Multi-target tracking in multiple non-overlapping cameras using constrained dominant sets. *arXiv preprint arXiv:1706.06196*, 2017. 7

[33] R. R. Varior, M. Haloi, and G. Wang. Gated siamese convolutional neural network architecture for human re-identification. In *ECCV*, 2016. 3

[34] F. Wang, W. Zuo, L. Lin, D. Zhang, and L. Zhang. Joint learning of single-image and cross-image representations for person re-identification. In *CVPR*, 2016. 3

[35] X. Wang and R. Zhao. Person re-identification: System design and evaluation overview. In *Person Re-Identification*, pages 351–370. 2014. 1, 3, 4, 6

[36] L. Wu, C. Shen, and A. v. d. Hengel. Deep recurrent convolutional networks for video-based person re-identification: An end-to-end approach. *arXiv preprint arXiv:1606.01609*, 2016. 3, 6, 7

[37] T. Xiao, H. Li, W. Ouyang, and X. Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *CVPR*, 2016. 3, 6

[38] F. Xiong, M. Gou, O. Camps, and M. Sznaier. Person re-identification using kernel-based metric learning methods. In *ECCV*. 2014. 3

[39] Y. Yan, B. Ni, Z. Song, C. Ma, Y. Yan, and X. Yang. Person re-identification via recurrent feature aggregation. In *European Conference on Computer Vision*, pages 701–716. Springer, 2016. 2, 3, 6, 7

[40] H. Yao, S. Zhang, D. Zhang, Y. Zhang, J. Li, Y. Wang, and Q. Tian. Large-scale person re-identification as retrieval. In *ICME*, 2017. 3

[41] H. Yao, S. Zhang, Y. Zhang, J. Li, and Q. Tian. Deep representation learning with part loss for person re-identification. *arXiv preprint arXiv:1707.00798*, 2017. 3, 6

[42] D. Yi, Z. Lei, and S. Z. Li. Deep metric learning for practical person re-identification. In *ICPR*, 2014. 3

[43] W. Zhang, S. Hu, and K. Liu. Learning compact appearance representation for video-based person re-identification. *arXiv preprint arXiv:1702.06294*, 2017. 7

[44] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *CVPR*, 2013. 3

[45] 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 *ECCV*, 2016. 1, 2, 3, 4, 6, 7

[46] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015. 1, 3

[47] Z. Zhou, Y. Huang, W. Wang, L. Wang, and 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 Conference on Computer Vision and Pattern Recognition*, pages 4747–4756, 2017. 7