Spatial-Temporal Attention-aware Learning for Video-based Person Re-identification

Guangyi Chen, Jiwen Lu, *Senior Member, IEEE,* Ming Yang, *Member, IEEE,* and Jie Zhou, *Senior Member, IEEE*

***Abstract*—In this paper, we present a spatial-temporal attention-aware learning (STAL) method for video-based person re-identification. Most existing person re-identification methods aggregate image features identically to represent persons, which are extracted from the same receptive field across video frames. However, the image quality may be varying for different spatial regions and changing over time, which shall contribute to person representation and matching adaptively. Our STAL method aims to attend to the salient parts of persons in videos jointly in both spatial and temporal domains. To achieve this, we slice the video into multiple spatial-temporal units which preserve the body structure of a person and develop a joint spatial-temporal attention model to learn the quality scores of these units. We evaluate the proposed method on three challenging datasets including iLIDS-VID, PRID-2011 and the large-scale MARS dataset, and consistently improve the rank-1 accuracy by a large margin of 5.7%, 0.9%, and 6.6% respectively, in comparison with the state-of-the-art methods.**

***Index Terms*—Person re-identification, Spatial-temporal attention model**

I. Introduction

Person re-identification (ReID) attempts to match pedestrians across multiple cameras, with great potential in surveillance applications [1]. It is such an intriguing vision problem because of complicated intra-camera variances in pose, illumination, viewpoint, partial occlusion, and cluttered background. Conventional works [2]–[5] address above problems by extracting robust invariant features and learning a discriminative metric subspace to accommodate inter-camera variances. However, the hand-craft features require strong prior knowledge yet lack of semantic clue. Recently, deep neural networks have been successfully applied for image-based person ReID [3], [6], [7] to learn discriminative image representations in an end-to-end fashion.

 This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFA0700802, in part by the National Natural Science Foundation of China under Grant 61822603, U1813218, U1713214, Grant 61672306, Grant 61572271, in part by the Shenzhen Fundamental Research Fund (Subject Arrangement) under Grant JCYJ20170412170602564, and in part by the internship of the first author at Horizon Robotics, Inc.(*Corresponding author: Jiwen Lu*)

Guangyi Chen, Jiwen Lu and Jie Zhou are with the State Key Lab of Intelligent Technologies and Systems, Beijing Research Center for Information Science and Technology (BNRist), Department of Automation, Tsinghua University, Beijing, 100084, China. E-mail: chen-gy16@mails.tsinghua.edu.cn; lujiwen@tsinghua.edu.cn; jzhou@tsinghua.edu.cn.

Ming Yang is with Horizon Robotics, Inc., Beijing & Shenzhen, 100080, China. E-mail: ming.yang@horizon-robotics.com.

EDICS: ARS-IVA: Image & Video Mid Level Analysis.

Person re-identification from videos captured by multiple cameras is a more practical setting than from still images and gains increasing research interests [8]–[19] recently. In fact, surveillance videos with pedestrians are the original data for image-based person ReID before pre-processing and preserve abundant and potentially complementary spatial-temporal characteristics of pedestrian, from different poses and view angles. However, identifying discriminative regions on pedestrian against distractions and aggregating their features are not straightforward for representing and matching persons in spatial-temporal domains. Naturally, most video-based ReID methods [20]–[22] employ a CNN-RNN structure to extract image features and apply average pooling to aggregate them. Nonetheless, in these ways, the matching of persons is sensitive to or may be misled by some "bad" samples due to occlusions or clutter background, since features from all frames and regions contribute equally to the matching. For example, when two persons are occluded by the same object, the similar appearance on occluded parts may result in a false positive in person ReID.

To address the above problems, we propose in this paper a spatial-temporal attention-aware learning (STAL) method for video-based person ReID. Motivated by the observations that image qualities in videos are varying in both spatial and temporal domains, the STAL method aims to jointly identify and match the spatial and temporal salient parts dynamically. As shown in Fig. 1, the proposed method avoids the misleading from "bad" parts of video by the attention mechanism. To achieve this, we slice the pedestrian video into multiple spatial-temporal units and develop a joint spatial-temporal attention model for evaluating the quality scores of individual units. The temporal attention model strives to identify those frames with clear and complete human figures in a representative gait. While the spatial attention model explores the discriminative salient body parts.

Specifically, our STAL framework includes three main branches: a *global representation branch*, a *local representation branch*, and a *spatial-temporal attention branch*. The global representation branch extracts perceptual features for the appearance of the entire body with a Convolutional Neural Network (CNN). The local representation branch crops different body parts with a pose estimation algorithm and collects local features of different body parts. Then, with the attention scores generated by the spatial-temporal attention branch, the local representations of different spatial-temporal units are aggregated into the final representation, where the possible blurred, ambiguous, or occluded units are down-weighted. To leverage the merits of complementary global feature and local

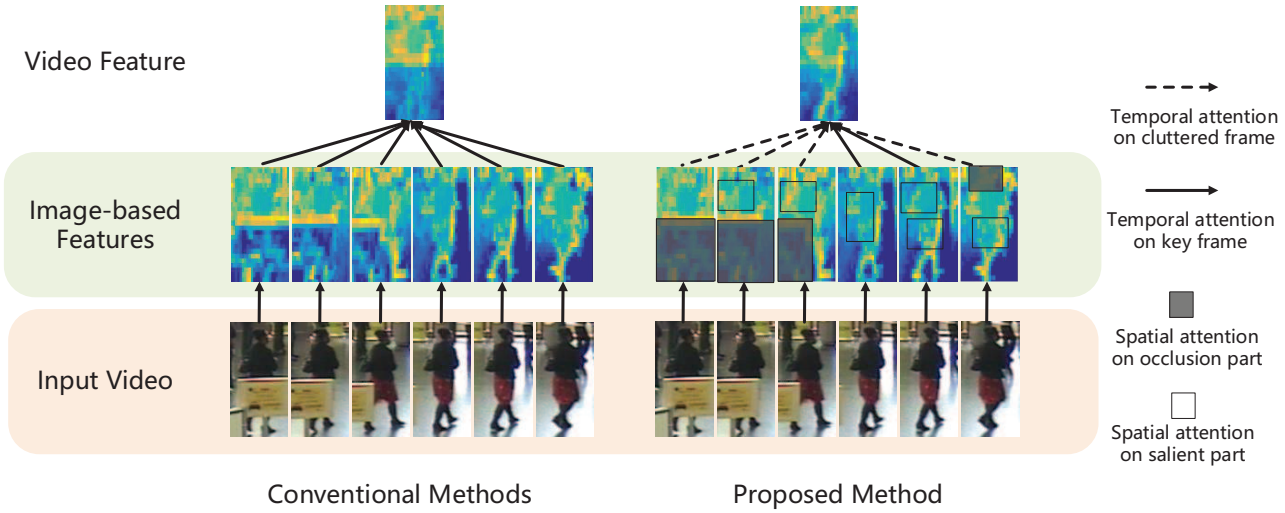


Fig. 1. Difference between proposed method and conventional methods. The left part shows that conventional methods treat image-based feature maps equally and aggregate them by average pooling. The right part shows that the proposed method considers spatial attention on image-based features and temporal attention for feature aggregation.(Best viewed in color)

feature, we integrate them in the metrics and construct an end-to-end deep neural network architecture to learn them simultaneously. Extensive experimental results on three public video datasets including PRID-2011 [23], iLIDS-VID [8], and MARS [12] demonstrate consistent improvement and superior cross-dataset generalization ability of our method.

We summarize three key contributions of our work as follows:

1) We propose a STAL algorithm to attend to the salient parts of pedestrian videos on both temporal and spatial dimensions. We slice the pedestrian video into multiple spatial-temporal units and focus on the "key" ones by learning a joint spatial-temporal attention score map.

2) We embed the spatial-temporal attention branch and CNN representation learning branch in an end-to-end framework and train it with a designed consistent loss.

3) We conduct extensive experiments on three challenging video datasets to demonstrate the efficacy of our STAL method. The results show that the proposed method outperforms other state-of-the-art methods.

II. Related Work

Recent years have witnessed the extensive studies of person re-identification. Existing ReID methods are roughly classified into two categories: *image-based approaches* and *video-based approaches*. In this section, we first review some related works about both image-based and video-based person ReID. Then, we briefly describe the attention model, which is widely applied in the person re-identification problem.

A. Image-based Person Re-identification

Previous image-based works [2], [4], [6], [18], [24]–[30] aim to extract the robust discriminative feature representation and learn effective metric distance. Feature learning methods [2], [24], [31]–[35] try to build a distinctive and robust image representation which is invariant to environmental and viewpoint changes. For example, LOMO [2] and GOG [24] are hand-crafted descriptors that combine the color and texture features. Salience match [32] and mid-level filter [34] find the salience patch of the pedestrian by a learning algorithm. Besides robust features, metric learning also plays a determining role in person ReID. Some methods [2], [36]–[39] learn a discriminant subspace or an integrated metric to emphasize inter-person distance and deemphasize intra-person distance. While Prosser et al. [40] formulate the person re-identification problem as a ranking problem and apply the RankSVM to learn a subspace for similarity measure. Recently, deep neural networks have been applied to person ReID successfully to jointly learn feature representation and similarity metric. Some works [3], [41], [42] formulate ReID as a binary verification problem and apply the siamese network to extract deep features. In addition, to preserve the rank relationship with a margin among a triplet of person samples, the triplet loss [1], [25], [30], [43] are proposed to learn robust CNN features. In addition, some other works [6], [44], [45] consider ReID as a multi-class recognition problem and learn the deep discriminative features with the softmax loss.

B. Video-based Person Re-identification

Video sequences provide abundant and diverse person samples and their possible correspondences in consecutive frames. Thus, video-based person ReID methods [8]–[19] take great efforts to leverage the motion cue and identify informative samples along the time axis for re-identification, ranging from low-level spatial-temporal feature extraction to high-level key frame selection or image-based feature aggregation. For example, some previous methods [9], [10], [14] take motion into consideration and employ the HOG3D features [46] on person videos as spatial-temporal features. Wang *et al.* [47] slice videos into multiple segments according to walking cycle, and select and rank discriminative frame segments. Then, to aggregate image-based features temporally, McLaughlin *et al.* [15] and Zhou *et al.* [19] apply a temporal pooling layer

to combine features captured by the CNN-RNN model; and QAN [48] designs a quality-aware network to estimate the image quality scores and integrates the frame-wise features weighted by these scores. While RFANet [49] employs a long short-term memory (LSTM) network to aggregate the image-based feature. Our work is related to QAN [48] in that we also integrate person features from an image sequence. A major difference is that we further develop a spatial attention model to find the salient regions on the person body, instead of only focusing on the frame-level attention [48].

C. Attention Model

Attention model [50] is a natural imitation of the human perception process which concentrates on what we are interested in. The visual attention mechanism has two common models: *recurrent attention perception* and *interested region mask*. Recently, many works [51]–[54] have implemented these two attention models with deep neural networks, especially using recurrent neural networks (RNN) and long short term memory (LSTM), for vision problems. For instance, Kelvin *et al.* [52] propose a recurrent attention model to learn a sequence of image attentions about each word in a caption. Xiao *et al.* [53] propose a two-level attention model to generate candidate patches and localize discriminative parts spatially. Attention models have been also applied for person ReID to learn the salient parts of persons to boost performance [19], [25]. Liu *et al.* [25] propose an end-to-end comparative attention network which designs an LSTM network to obtain multiple attention maps. Zhou *et al.* [19] employ an attentive temporal RNN model to represent videos and apply a spatial recurrent model on pair-wise metric learning. Different from these attention methods, our STAL method learns a spatial-temporal attention score map to indicate the qualities of different parts of the pedestrian.

III. Approach

The goal of our STAL method is to jointly explore the salience of person videos in both spatial and temporal domains. Therefore, we first slice the videos into multiple spatial-temporal units and then learn an attention score map which indicates the qualities of all space-time bins.

A. Overall Architecture

Given a pedestrian video $X = \{x^t\}_{t=1:T}$, where T is the number of video frames and x^t denotes the tth frame. As shown in Fig. 2, the proposed network is divided into two branches: a global representation branch and a local representation branch. They are fused in an end-to-end framework to learn discriminative person representation in different granularities.

The global representation branch is designed to learn the full body representation of pedestrians. In this branch, at the beginning of the process, image sequence X is fed into a low-level CNN to generate the low-level representations, after that, we apply a residual attention network (RAN) in a high-level CNN to extract global features $g^t = \mathcal{G}(x^t)$. Please refer to section 3.2 for the details of RAN.

The local representation branch is used to address the local variance of person video, *e.g.,* local mismatching due to pose variance. In this branch, we first apply a human pose estimation algorithm [28] to locate the body joints and generate the body part coordinates $p^{r,t} = \{p_1, p_2, p_3, p_4\}^{r,t}$ based on the estimated joints. The local part generator is pretrained with the MPII human pose dataset [55]. Then we apply an ROI pooling layer with body part coordinates on the feature maps by the low-level CNN and design a part-specific network (the light blue box in Fig.2) to generate the local part representation. The part-specific network has the same structure for different body parts but learns the different parameters. The whole processing is formulated as:

$$\{f^{r,t}\}_{r=1:R,t=1:T} = \mathcal{F}(x^t, p^{r,t}), \tag{1}$$

where $f^{r,t}$ represents the local feature of rth spatial body part in X^t.

In addition, we also develop a separate attention branch to learn a joint spatial-temporal attention score map $a^{r,t} = \mathcal{A}(X)$, which is used to evaluate the qualities of different spatial-temporal units. The temporal attention focuses on the key frames with rich discriminative information, while the spatial attention identifies the body parts which are not corrupted by occlusions and cluttered background. Finally, we define an aggregation function of local features $f^{r,t}$ with the attention scores $a^{r,t}$ to calculate the distance between two pedestrian video clips, which is formulated as:

$$d_l(i,j) = \psi(f_i^{r,t}, f_j^{r,t}; a_i^{r,t}, a_j^{r,t}), \tag{2}$$

where i, j denote two person videos captured in different cameras. We aggregate the global features with different frames in a temporal pooling layer and calculate the global distance as $d_g(i,j) = ||g_i - g_j||_2$, where g_i denotes the global feature of ith person. In the training procedure, we optimize the objective function with both global and local distance, while in the testing procedure, we add them for the final similarity measure.

The objective function of our method is formulated as follows:

$$\min_{\mathcal{G},\mathcal{F},\mathcal{A}} = L_{tri}(\mathcal{G},\mathcal{F},\mathcal{A}) + L_{cls}(\mathcal{G}) + C_{cons}(\mathcal{F}) \tag{3}$$

which contains three parts: triplet loss, softmax loss, and consistency constraint.

1) Triplet Loss: We design the triplet loss to preserve the rank relationship among a triplet of pedestrian videos. In the triplet loss, the distances between feature pairs from the same class are minimized while the ones from different classes are maximized. We calculate the triplet loss with both global and local features as follows:

$$\begin{aligned} L_{tri} = &\sum_{i,j,k\in\Omega} [d_g(i,j) - d_g(i,k) + m_g]_+ \\ &+ \sum_{i,j,k\in\Omega} \lambda[d_l(i,j) - d_l(i,k) + m_l]_+, \end{aligned} \tag{4}$$

where m_g and m_l are margin thresholds to limit the gap between the distances from positive and negative samples, and $[x]_+$ denotes the max function $max(0,x)$.

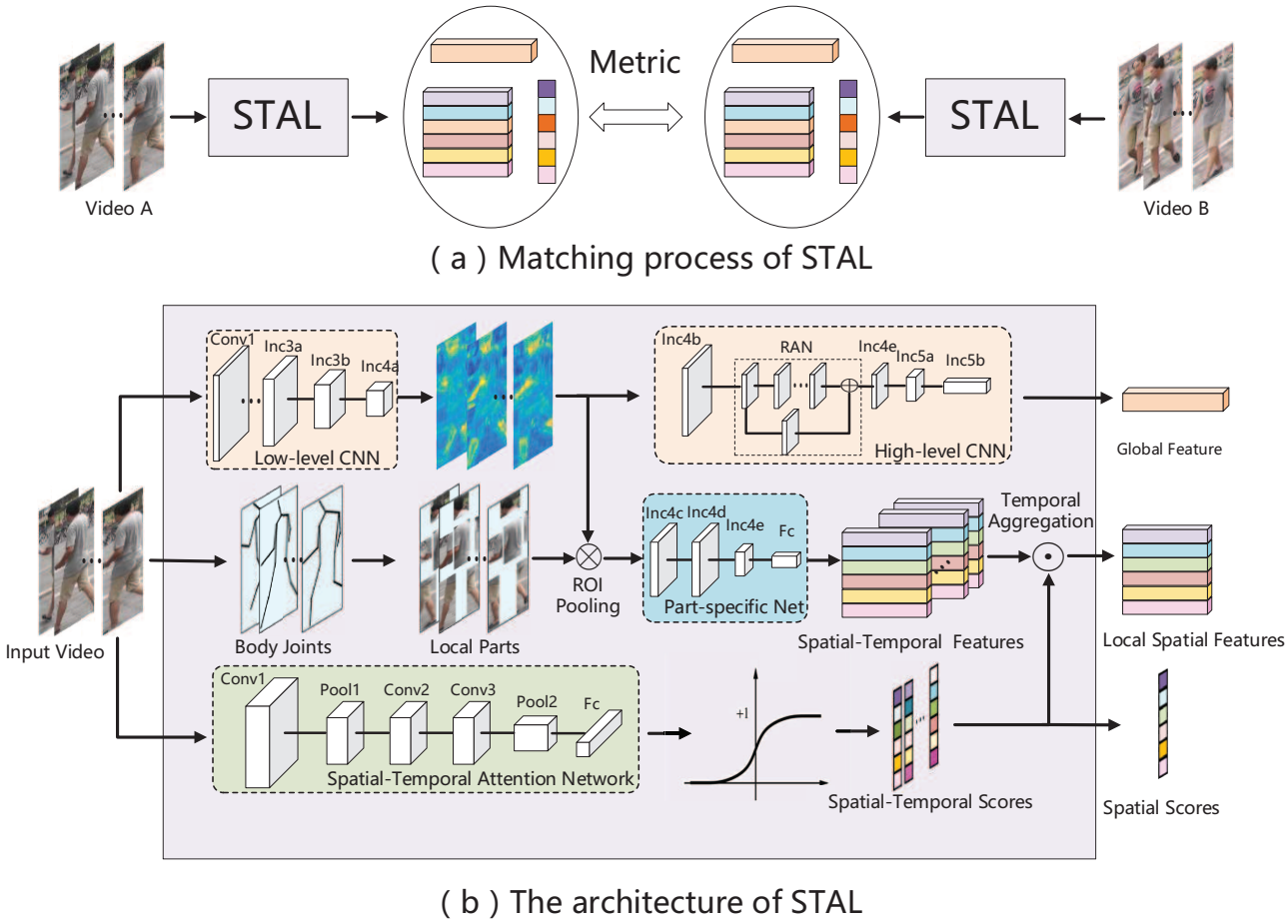


Fig. 2. The network architecture of the STAL algorithm. (a) illustrates the procedure that we calculate the metric between two person videos by the global feature, local features and spatial attention scores learned by STAL. Specifically, we measure the global distance with global feature and calculate the local distance with local features of different spatial body parts and corresponding spatial attention scores. Then we add the global and local distances for similarity measure. (b) shows the detailed architecture of our STAL method, which extracts the global feature, local features and spatial attention scores of the given video clip. The network has three branches: a global representation branch, a local representation branch, and a spatial-temporal attention branch. In the global representation branch, we apply a CNN to capture the overall appearance information of each frame and aggregate these features with temporal pooling. In the backbone CNN, we build an additive residual attention network (RAN) to highlight the human body. For the local representation branch, we first apply a human pose estimation algorithm to locate the body joints and generate bounding boxes of different body parts. Then we use the ROI pooling layer and part-specific networks to generate local representations. To effectively aggregate these local representations of spatial-temporal parts, we design a spatial-temporal attention branch to learn the attention scores of individual frames and body parts as weighting coefficients. (Best viewed in color)

2) Softmax Loss: We apply the softmax loss function to learn the identity-specific global representation. Different from triplet loss, the softmax loss focuses on the robustness of person video representations for identification. The softmax loss is defined as:

$$L_{cls} = \sum_{i \in \Omega} \frac{exp(W^g_{y_i} g_i)}{\sum_k exp(W^g_k g_i)}, \tag{5}$$

where y_i is the identity of ith person and $W^g_{y_i}, W^g_k$ indicate y_ith and kth columns of the softmax matrix.

3) Consistency Constraint: To preserve the consistency between local and global features, we develop a consistency constraint in our framework, which requests that the identifies of single local feature and global feature are identical:

$$C_{consis} = \sum_{i \in \Omega} \sum_{r,t} \frac{exp(W^r_{y_i} f^{r,t}_i)}{\sum_k exp(W^r_k f^{r,t}_i)}, \tag{6}$$

Note that the same parts in different frames share the same softmax matrix.

B. Spatial-Temporal Feature Learning

To capture the overall appearance of pedestrians, we learn global features on the person's whole body. To cope with the variance of person pose and leverage the body structure information, we propose to slice the pedestrian videos into multiple spatial-temporal units and extract local features. As shown in Fig. 2, the low-level CNN is shared for both local features and global feature, which are applied to collect semantic information. The residual attention structure is applied to extract the global features, while a body part generation model and an ROI pooling layer are employed to learn local features.

Residual Attention Network Inspired by [56], we build a residual attention branch on the CNN backbone to highlight the human body. Instead of the stack architecture, we only add the attention branch on the low-level feature maps which have

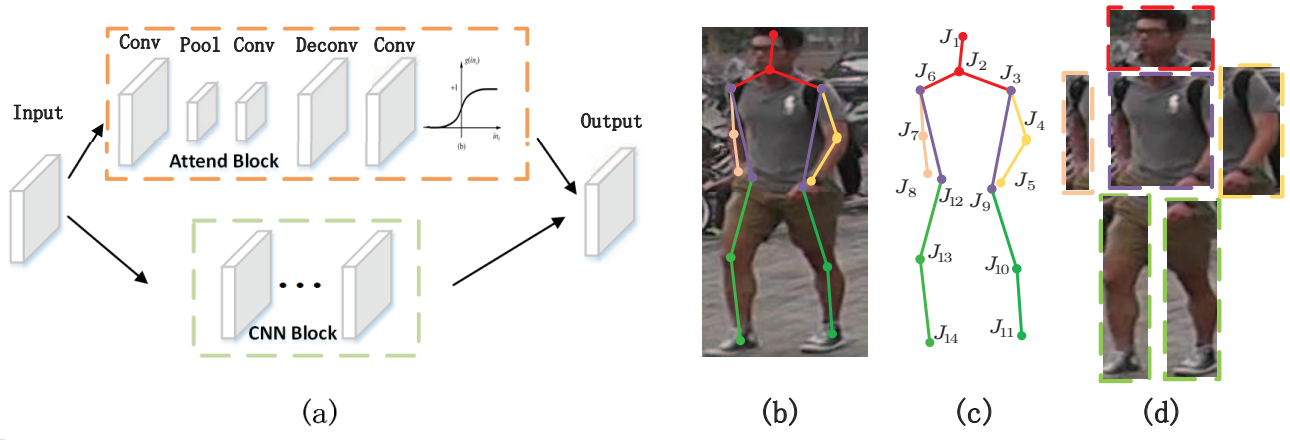


Fig. 3. Illustration of Residual Attention Network and Local Parts Generation. (a) The detail architecture of Residual Attention Network. (b) Body joints of a person image in MARS. (c) Body joints assignment and local part generation. (d) A example of generated local body parts.

enough resolution to estimate the attention, since the higher level feature maps are too coarse to observe precise attention masks. As shown in part (a) of Fig. 3, the residual attention branch is divided into a downpooling module, an uppooling module and a sigmoid layer. To lower computational cost, we further sample the downpooling module and uppooling module in the [56]. Specifically, the downpooling module contains 2 convolution layers and a max pooling layer to extract the discriminative information of input feature maps. While the uppooling module, which contains a deconvolution layer and a convolution layer, recovers the resolution of the mask to raw input. We also apply a sigmoid layer to normalize the attention mask range to $[0, 1]$. In the end, we multiple the soft masks with the original outputs of the CNN network as the output of the attention branch and sum up the outputs of attention block and CNN block.

Local Parts Generation In the temporal dimension, the pedestrian video is sliced into frames. We assume that frames are sampled at regular intervals and take every frame as a temporal unit. While in the spatial dimension, we consider about human body structure and segment a person crop into multiple body part regions.

Given a person crop, we first locate 14 joints of human body $J_i \in R^{X \times Y}{}_{i=1:14}$ as Fig. 3 (b) with a human pose estimation network [28] which is pretrained on the MPII human pose dataset [55]. We incorporate the pose estimation network into our end-to-end framework and refine it with the ReID training set. Then, we utilize these 14 joints to generate 6 coarse body part regions, which corresponds to head, torso, right arm, left arm, right leg and left leg. As shown in Fig. 3 (c), we assign the 14 body joints to these 6 parts, the head part $P_1 = [J_1, J_2, J_3, J_6]$, the torso part $P_2 = [J_2, J_3, J_6, J_9, J_{12}]$, the right arm part $P_3 = [J_3, J_4, J_5]$, the left arm part $P_4 = [J_6, J_7, J_8]$, the right leg part $P_5 = [J_{12}, J_{13}, J_{14}]$, and the right leg part $P_6 = [J_9, J_{10}, J_{11}]$, respectively. The bounding box of corresponding part is generated by joints as:

$$\begin{aligned} p^r &= \{p_1, p_2, p_3, p_4\}^r \\ &= \{\min_{J \in P_r} x(J), \max_{J \in P_r} x(J), \min_{J \in P_r} y(J), \max_{J \in P_r} y(J)\}^r \end{aligned} \quad (7)$$

An example of the generated local body parts is visualized in Fig. 3 (c).

As shown in Fig. 2, given the generated local body parts p^r, and the feature maps by the low-level CNN, we apply an ROI pooling layer to obtain the local feature maps of different body parts. Then we additionally apply a part-specific CNN network to generate the feature representation of local body parts based on the local feature maps. The part-specific networks of different body parts are independent with the same architecture. The local part estimator is applied softly in our STAL method and refined with the person ReID objective. This is being said, the estimated body part regions with deviations are not sensitive for the person verification since its impact is relatively down-weighted by the following attention model. This scheme makes it flexible to adopt any pre-trained pose estimation model.

C. Spatial-Temporal Attention Model

In the pedestrian videos of the real surveillance system, the qualities of different frames and regions vary dramatically due to occlusions, illumination and cluttered background. Some "bad" samples may mislead the matching of pedestrian videos if we treat all samples equally. To promote the discriminative samples and punish the "bad" ones, we propose a spatial-temporal attention network (STAN), for learning the attention scores of individual frames and body parts. Fig. 2 (b) shows the detailed network architecture of the STAN, which includes a 7×7 convolution layer, two 3×3 convolution layers, a 7×7 max pooling layer, a fully connected layer whose output is $T \times R$ and a sigmoid layer. Three convolution layers are applied to gather image textural information and the fully connected layer generates the raw attention scores. To normalize the scores into $[0, 1]$, we employ a sigmoid layer on the attention scores. Finally, we apply the $L1$ normalization layer for the attention map to avoid the mismatch caused by scale offset. The output of STAN, $\{a^{r,t}\}$, is a $T \times R$ score map, which corresponds to T frames and R body parts.

Then, with the attention score map $a^{r,t}$, we aggregate all local spatial-temporal features $f^{r,t}$ on both temporal and

spatial domains in an attention-aware way. To promote the high-quality spatial-temporal units and down-weight the low-quality ones in the metric, we denote the aggregated function $\psi(\cdot)$ as:

$$\psi(f_i^{r,t}, f_j^{r,t}; a_i^{r,t}, a_j^{r,t}) = \sum_{r=1}^{R} \frac{\bar{a}_i^r \bar{a}_j^r}{||\bar{a}_i||||\bar{a}_j||} ||f_i^r - f_j^r||_2, \quad (8)$$

where f_i^r denotes the spatial representation aggregated on the temporal dimension, which is formulated as:

$$f_i^r = \frac{\sum_{t=1}^{T} a_i^{r,t} f_i^{r,t}}{\sum_{t=1}^{T} a_i^{r,t}}. \quad (9)$$

$\bar{a}_i^r$ is the attention score of spatial feature f_i^r, which is evaluated by the mean of score map on the temporal dimension. $||\bar{a}_i||$ is a normalization term on the spatial scores to avoid the scale offset.

Guided by the spatial-temporal attention scores, the part pairs with higher scores lead the matching of two individuals. Finally, we aggregate the local features of all spatial-temporal parts to form a fixed feature vector to represent videos with a variable number of frames. Note that the average pooling is a special case of the proposed attention-aware aggregated method when all attention scores are identical in both temporal and spatial dimensions.

D. Backpropagation

In this subsection, we calculate the backpropagation of attention scores in two stages: metric calculation and temporal feature aggregation. For simplification, we only take the gradients of the attention score of one person $a_i^{r,t}$ into account.

In the metric calculation stage, the loss of metric ψ back propagates to the regional video representations f_i^r and normalized regional attention scores ${a_{n_i}}^r = \frac{\bar{a}_i^r}{||\bar{a}_i||}$, which is formulated as follows:

$$\begin{aligned} \frac{\partial \psi}{\partial f_i^r} &= {a_{n_i}}^r {a_{n_j}}^r \frac{f_i^r - f_j^r}{||f_i^r - f_j^r||_2} \\ \frac{\partial \psi}{\partial {a_{n_i}}^r} &= \sum_{r=1}^{R} {a_{n_j}}^r ||f_i^r - f_j^r||_2. \end{aligned} \quad (10)$$

While in the temporal feature aggregation stage, the backpropagation of f_i^r is formulated as:

$$\frac{\partial f_i^r}{\partial a_i^{r,t}} = \frac{f_i^{r,t} - f_i^r}{\sum_{t=1}^{T} a_i^{r,t}}. \quad (11)$$

The gradient of regional attention scores ${a_{n_i}}^r$ as Eq. 6:

$$\frac{\partial {a_{n_i}}^r}{\partial a_i^{r,t}} = \frac{1 - {a_{n_i}}^r}{||\bar{a}_i||}. \quad (12)$$

Thus we formulate the propagation process of ψ as:

$$\frac{\partial \psi}{\partial a_i^{r,t}} = \frac{\partial \psi}{\partial f_i^r} \frac{\partial f_i^r}{\partial a_i^{r,t}} + \frac{\partial \psi}{\partial {a_{n_i}}^r} \frac{\partial {a_{n_i}}^r}{\partial a_i^{r,t}} \quad (13)$$

TABLE I
DETAILED STRUCTURE OF OUR PROPOSED CNN

layer name	kernel size	stride	output size
input			$3 \times 224 \times 224$
conv1	7×7	2	$64 \times 224 \times 224$
pool1	3×3	2	$64 \times 112 \times 112$
conv2_1 $\times$ 1	1×1	1	$64 \times 112 \times 112$
conv2_3 $\times$ 3	3×3	1	$192 \times 112 \times 112$
pool2	3×3	2	$192 \times 56 \times 56$
inception3a		2	$256 \times 28 \times 28$
inception3b		1	$320 \times 28 \times 28$
inception3c		2	$576 \times 14 \times 14$
inception4a		1	$576 \times 14 \times 14$
inception4b		1	$576 \times 14 \times 14$
inception4c		1	$576 \times 14 \times 14$
ran_conv1	1×1	1	$64 \times 14 \times 14$
ran_pool1	3×3	2	$64 \times 7 \times 7$
ran_conv2	3×3	1	$192 \times 7 \times 7$
deconv	4×4	2	$192 \times 14 \times 14$
conv3	1×1	1	$576 \times 14 \times 14$
inception4d		1	$576 \times 14 \times 14$
inception4e		2	$1024 \times 7 \times 7$
inception5a		1	$1024 \times 7 \times 7$
inception5b		1	$1024 \times 7 \times 7$
pool_f	7×7	1	1024
inception4c_part		1	$576 \times 8 \times 8$
inception4d_part		1	$576 \times 8 \times 8$
inception4e_part		2	$1024 \times 4 \times 4$
fc_part			1024

E. Implementation Details

We select Caffe [57] as the basic toolbox to implement our experiments. To make the network easy to converge, we apply the GoogleNet which is pretrained on the ImageNet as the basic backbone network. Specifically, we follow the network definition and implementation details of the backbone GoogleNet of Caffe as *https://github.com/BVLC/caffe/tree/master/models/bvlc_googlenet*. As shown in Fig. 2 (b), we split the GoogleNet into the low-level CNN and high-level CNN at the inception-4b layer, since the inception-4c layer has a better trade-off between semantic information and resolution. As shown in Table. I, we introduce the structure of our CNN in the detail by listing the kernel size, stride and output size of each layer. Note that we design the part-specific network for different body parts with the same structure as $inception\_part$. The output of low-level CNN and part coordinates are fed into the ROI pooling layer to generate part representations.

In the training stage, we feed a triplet of person videos into our network as a batch, which contains an anchor sample, a positive sample, and a negative sample. We sample $T = 8$ frames of a video into a batch and resize them to 224×224. The middle representations generated by the shared low-level CNN are 14×14 feature maps. We apply the ROI pooling layer on them to get 8×8 local middle representation of different body parts. Then, the three inception layers and an FC layer are employed to generate local features. Finally, the global representation is a 1024 dimension feature pooled by GoogleNet, while local representations are also 1024 dimension features generated by FC layer. We set global margin $m_g = 1.2$ and local margin $m_l = 0.8$ in the Eq. 4. To have a trade-off, we

TABLE II
THE BASIC INFORMATION OF ALL DATASETS IN THE EXPERIMENTS

Datasets	PRID-2011	iLIDS-VID	MARS
Identities	200	300	1261
Tracklets	400	600	21K
Cameras	2	2	6
Images	42K	44K	1.1M
Crop Size	128×64	Vary	256×128
Label Method	Hand	Hand	DPM+GMMCP
Splits	Random	Random	Fixed
Matching	Closed-Set	Closed-Set	Open-Set
Evaluation	CMC	CMC	CMC & mAP

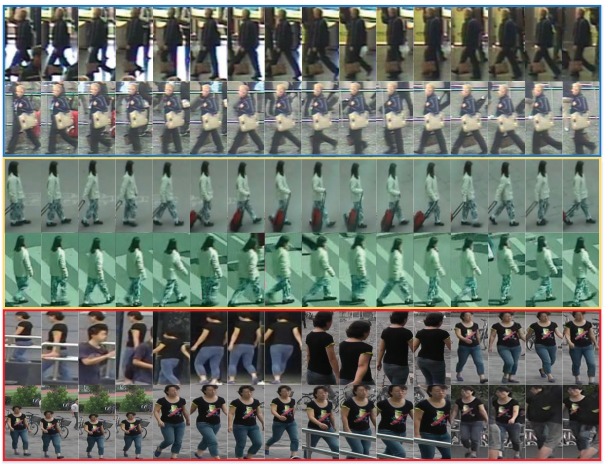


Fig. 4. Samples of iLIDS-VID, PRID-2011 and MARS datasets. The top two rows show the video from PRID-2011, the middle two rows are sampled from iLIDS-VID, and the bottom part is the pedestrian samples from MARS.

set the rate between global triplet loss and local triplet loss in the Eq. 4 as $\lambda = 0.3$. We apply the SGD as the optimizer with momentum 0.9 and gamma 0.5. The learning rate is set to 0.001 and the weight decay factor for L2 regularization is set to 0.0002. We train our model for 13500, 6500 and 150000 iterations(triplets) on the iLIDS-VID, PRID-2011 and MARS dataset respectively. Note that, we do not adopt any data argumentation methods (e.g., scaling, rotation, flipping, and color distortion)

During the evaluation, we first extract both global features and local features and aggregate them with learned temporal attention score. Then we calculate the similarity matrix for each feature independently with Euclidean distance and weight all local distance matrices by spatial attention scores. Finally the global distance and local distance are fused by a rate as $d = d_l + 0.3d_g$.

IV. EXPERIMENTS

We evaluated our method on three available pedestrian video datasets including iLIDS-VID [8], PRID-2011 [23], and MARS [12] and compared with the baseline methods and other state-of-the-art methods.

A. Datasets and Settings

The basic information of three datasets in our experiments is summarized in Table II and some pedestrian video samples are displayed in Fig. 4.

1) PRID-2011: The PRID-2011 dataset contains videos recorded from two non-overlapping surveillance cameras. 385 persons are under camera A, while 749 persons are under camera B. Among all pedestrian videos, 400 videos of 200 pedestrians are captured in both two cameras. Each video has 5 to 675 frames with an average number of 100. To ensure the effective length of videos, we selected videos of 178 identities with more than 27 frames. As shown in the top of Fig. 4, this dataset was captured in uncrowded outdoor scenes with large illumination and viewpoint change.

2) iLIDS-VID: The iLIDS-VID dataset contains 600 pieces of videos for 300 randomly sampled people. Each person has one pair of video from two camera views, which have variable lengths from 23 to 193 frames with an average of 73 frames. As shown in the bottom of Fig. 4, the videos captured in a crowded airport arrival hall are challenging for similar appearance, cluttered background, and random occlusion.

3) MARS: The Motion Analysis and Re-identification Set (MARS) is a video extension of Market1051 [58] dataset which contains 1261 persons and around 20000 video sequences. These sequences are captured by 6 cameras at most and 2 cameras at least, from which each identity has 13.2 sequences on average. Videos of MARS dataset generated by a DPM detector [59] and a GMMCP tracker [60], instead of hand-drawn bounding boxes. The challenge of this dataset is largely in the variance of viewpoints and complicated occlusions.

Following the protocol of [8] for PRID-2011 and iLIDS-VID datasets, we labeled videos from the first camera as the probe set, while the others as the gallery set. We randomly split persons into equal-sized training and testing sets and repeated experiments 10 times and calculated the average accuracy. The experimental setup of MARS is the same as [12], which fixes the partition of 625 train persons and 634 test persons. We resorted to both cumulative matching characteristic (CMC) curve and mean Average Precision (mAP) as the evaluation metrics. CMC curves record the true matching within the top n ranks, while mAP scores consider precision and recall to evaluate the overall performance of methods. Noting that mAP needs to calculate recall for multiple ground truths, we only calculated CMC curve on PRID-2011 and iLIDS-VID datasets which only have one ground truth.

B. Comparison with the State-of-the-Art Methods

In this subsection, we compared the proposed approach with other state-of-the-art methods on three challenging datasets which include DVDL [10], DVR [8], TDL [14], STFV3D [11], RFANet+RSVM [49], LMKDCCA [61], CNN+RNN [20], CNN+XQDA [12], BRNN [22], GRU [7], AMOC [21], TAM+SRM [19], QAN [48], RQEN [63], TriNet [43], and ASTPN [62].

iLIDS-VID: Table III shows the performances of our STAL approach and most existing video-based person re-identification approaches on iLIDS-VID. Our STAL method improves over all of the compared methods and outperforms the second best method substantially by 6%, due to the improvement of the spatial-temporal salience parts learning.

TABLE III
COMPARISON WITH STATE-OF-THE-ART PERSON RE-IDENTIFICATION METHODS ON THE iLIDS-VID DATASET.

Datasets	iLIDS-VID			
Rank@R	R=1	R=5	R=10	R=20
DVDL [10]	25.9	48.2	57.3	68.9
DVR [8]	41.3	63.5	72.7	83.1
TDL [14]	56.2	88.2	95.3	97.8
STFV3D+KISSME [11]	44.3	71.7	83.7	91.7
RFANet+RSVM [49]	49.3	76.8	85.3	90.0
LMKDCCA [61]	73.3	90.5	94.7	98.1
CNN+RNN [15]	58.0	84.0	91.0	96.0
CNN+XQDA(MARS) [12]	54.1	80.7	90.0	95.4
BRNN [22]	55.3	85.0	91.7	95.1
GRU [7]	49.8	77.4	90.7	94.6
AMOC+ EpicFlow [21]	68.7	94.3	**98.3**	99.3
TAM+SRM [19]	55.2	86.5	-	97.0
QAN [48]	68.0	86.6	95.4	97.4
ASTPN [62]	62.0	86.0	94.0	98.0
RQEN [63]	77.1	93.2	97.7	**99.4**
STAL	**82.8**	**95.3**	97.7	98.8
Baseline	76.7	93.5	96.4	98.5

TABLE IV
COMPARISON WITH STATE-OF-THE-ART PERSON RE-IDENTIFICATION METHODS ON THE PRID-2011 DATASET.

Datasets	PRID-2011			
Rank@R	R=1	R=5	R=10	R=20
DVDL [10]	40.6	69.7	77.8	85.6
DVR [8]	48.3	74.9	87.3	94.4
TDL [14]	58.6	80.8	87.4	93.3
STFV3D+KISSME [11]	62.5	83.6	88.1	89.9
RFANet+RSVM [49]	58.2	85.8	93.4	97.9
LMKDCCA [61]	86.4	97.5	**99.6**	**100**
CNN+RNN [15]	70.0	90.0	95.0	97.0
CNN+XQDA [12]	77.2	93.1	96.7	99.1
BRNN [22]	72.8	92.0	95.1	97.6
GRU [7]	42.6	70.2	86.4	92.3
AMOC+ EpicFlow [21]	83.7	98.3	99.4	**100**
TAM+SRM [19]	79.4	94.4	-	99.3
QAN [48]	90.3	98.2	99.3	**100**
ASTPN [62]	77.0	95.0	99.0	99.0
RQEN [63]	91.8	98.4	99.3	99.8
STAL	**92.7**	**98.8**	99.5	**100**
Baseline	90.3	98.7	**99.6**	99.6

We obtained considerable improvement on the Rank 1, since the proposed method is insensitive to occlusions and cluttered background, which are major challenges in this dataset.

PRID-2011: As shown in Table IV, we also achieved the state-of-the-art performance on the PRID-2011 dataset and outperform 1% than the second best method. The improvement on PRID-2011 is less than the iLIDS-VID dataset since the differences among frames and regions are limited. The main challenge of PRID-2011 dataset is illumination variance between two camera views. Thus, the improvement of our STAL method on this dataset is not that significant.

MARS: MARS is a large-scale and realistic dataset since it was captured in a scene of the crowded supermarket with a complex environment. Different from the other two datasets, pedestrian videos of MARS are captured by six cameras. We evaluated our method without post-processings which are orthogonal to our method and could be integrated in a straightforward manner, such as various re-ranking schemes and data augmentation [6], [64], [65]. As illustrated in Table V, notably

TABLE V
COMPARISON WITH STATE-OF-THE-ART PERSON RE-IDENTIFICATION METHODS ON THE MARS DATASETS.

Datasets	MARS			
Rank@R	R=1	R=5	R=20	mAP
CNN+RNN [20]	56	69	73	77
CNN+XQDA [12]	65.2	82.4	89.0	48.0
TAM+SRM [19]	70.6	90.0	97.6	50.7
AMOC+ EpicFlow [21]	68.3	81.4	90.6	52.9
QAN [48]	72.1	85.5	93.2	50.2
ASTPN [62]	44	70	81	-
RQEN [63]	73.7	84.9	91.6	51.7
TriNet [43]	79.8	91.4	-	67.7
STAL + ResNet50	**82.2**	**92.8**	**98.0**	**73.5**
STAL	80.3	90.9	96.5	64.5
Baseline	71.5	83.3	89.9	50.8

STAL method also achieved the state-of-the-art performance on the Rank 1 accuracy on the MARS dataset. Different from iLIDS-VID dataset, the great performance on MARS indicates the effect of joint spatial-temporal learning and body-structure information to overcome the large variances with pose and view changes.

Comparison and Analysis: The QAN [48] method is a baseline method which only learns frame-level temporal quality scores and aggregates the features on temporal domain. Compared with QAN, our proposed method further consider the attention scores of human body parts and learn the joint spatial-temporal attention. The proposed method also has the advantage over TAM+SRM [19], which learns the spatial and temporal information without explicit attention mechanism. TAM employs RNN to learn temporal information and use temporal pooling to aggregate features of every frame. However, it is difficult to apply the RNN for pedestrian videos due to the large variance of the pose. The experiments on [14] have shown that the temporal features like walking actions or other motions have smaller inter-class variations and difficult to distinguish. ASTPN [62] designs a joint spatial-temporal pooling network to learn the spatial and temporal attention. However, the performance of ASTPN is limited by the simple attentive mechanism and weak feature representation, especially on the large-scale dataset, like MARS. Different from the ASTPN method which uses a matrix to learn the spatial-temporal attention, the proposed approach learns the intuitive attention scores of different body parts. RQEN [63] also considers the region-based spatial attention of the person image. However, the relation between the temporal dimension and spatial dimension is ignored. We introduce the body-structure information in our attention model. TriNet [43] is a general method for person re-identification, which is not specific to the video-based problem. In fact, TriNet is also a baseline method of our STAL method. In our method, we also apply the Triplet loss in Eq 4 to preserve the rank relationship among a triplet of samples, which is similar to TriNet. Moreover, we improve the TriNet with a spatial-temporal attention-aware learning method to model the video data and extract video-based feature representation and add the auxiliary Softmax supervisory signal. Compared with TriNet experimentally, STAL method has no advantage on both rank-5 and mAP scores, since the TriNet employ an advanced

ResNet50 as backbone network while our STAL applies GoogleNet in the previous manuscript for a fair comparison with baseline method QAN [48]. In addition, we implemented our STAL method with similar ResNet50 backbone network and obtained consistent improvement over TriNet on both rank-based accuracies and mAP scores.

C. Ablation Experiments

To show the effectiveness of our proposed STAL method, we conducted several ablation experiments about attention models; feature representation branches and local body parts.

1) Evaluations of Each Attention Model: To investigate the contribution of each attention component in the proposed method, we evaluated the baseline method without attention model, temporal attention network (TAN), spatial attention network (SAN) and proposed STAL method on iLIDS-VID datasets.

Baseline: For a fair evaluation of each attention component, we trained a baseline model without the modification of the network. While in the testing stage, we designed some switches to control the state of attention mechanisms. In the baseline method, we closed off all attention models by a matrix which is filled with ones, instead of the real spatial-temporal attention scores.

TAN: Compared with the baseline method, TAN opened the switch on temporal frame-level attention to aggregate the features of frames. Then, we calculated the distances of the aggregated features of different body parts respectively and summed them up equally as the final distance measure.

SAN: In contrast, SAN closed the switch on temporal attention and opened the one on spatial attention. In the temporal domain of SAN, we aggregated the features of frames by average pooling. We calculated the average of spatial-temporal attention scores along the temporal domain as the spatial attention scores. Then we weighted the distances of body parts by them.

STAL: STAL combined the TAN and SAN by opening all attention modules. We aggregated the features of different frames with temporal attention scores and calculate the final distance by the spatial attention scores.

Comparison and Analysis: Fig. 5 summarizes the performances of the different variants of the proposed method. It is easy to draw the following conclusions from the rank CMC curves of different variants.

1. By comparing Baseline and TAN, we concluded that the temporal attention model learns how discriminative of different frames in a video. By excluding noise frames in matching, we obtained more robust feature aggregation than average pooling.

2. The improvement between SAN and Baseline demonstrated that the spatial attention models have explored the salient body parts for each person.

3. When combining the TAN and SAN in a joint STAL model, we further improved the performance. It indicates that the components of the proposed method not only works separately but also can be combined to boost performance.

2) Evaluations of Each Feature Branch: We also designed an ablation experiment to analyze the contributions of each feature branch. First, we independently evaluated two basic feature branches including the pure global feature branch (PGB) and the pure local feature branch (PLB). In these two evaluations, only single feature branch is applied in both training and testing stages. Then we jointly trained both two feature branches and test each branch respectively, the global branch (GB) or the local branch (LB). To investigate the contribution of designed residual attention model (RAN), we also conducted an experiment where only the residual attention model is ablated.

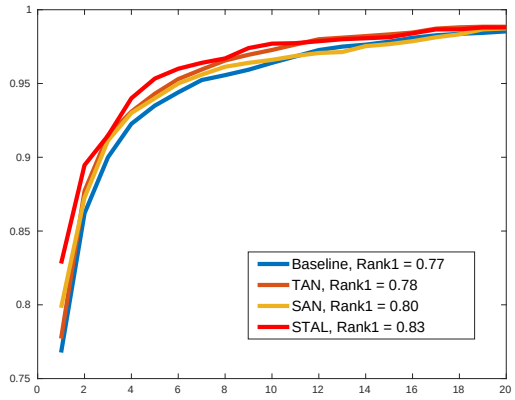


Fig. 5. Performance of variants of proposed method on the iLIDS-VID dataset. TAN refers to temporal attention network and SAN represents the network which only considers spatial attention. Finally STAL stands for joint spatial-temporal attention network.

We compared the CMC curves of the feature-branch ablation experiments on ILIDS-VID dataset in Fig. 6. First, the comparison between PGB and GB and the one between PLB and LB show that joint learning of global feature and local feature boosts the robustness of feature representation. When we learn the global feature, the local body parts learning provide structural information and local attention to improve feature representation. Meanwhile, the global feature learning process enhances the semantic expression ability of feature map, which assists to detect local body parts. Second, the performance of complete STAL advances than other single branch baselines, which indicates that STAL learns the complementary representation information between global feature and local feature. Third, the ablation experiment of residual attention model shows the important contribution of RAN in the proposed STAL method.

3) Evaluations of Each Local Part: To explore the salient spatial parts of person images, we segmented a whole person image into different body parts and learned corresponding representations and attention scores. In experiments, we naturally segmented images into the head part, torso part, right arm part, left arm part, right leg part and left leg part. We conducted experimental analysis on the iLIDS-VID dataset to show the efficiencies of different local parts. Specifically, to avoid the interference of global branch, we only tested the local branch of STAL. In the testing stage, we extracted local features of different body parts and respectively calculated the Euclidean distance to measure similarity.

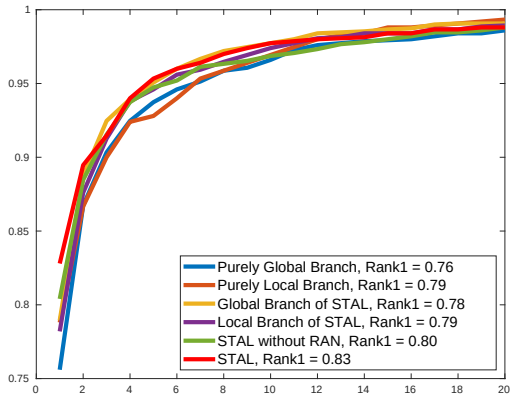


Fig. 6. CMC curves of proposed algorithm with different body parts on the iLIDS-VID dataset.

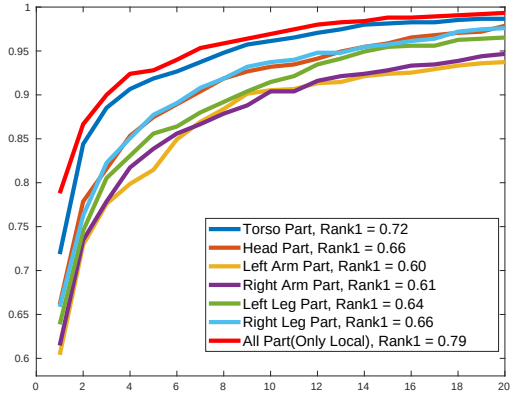


Fig. 7. CMC curves of proposed algorithm with different body parts on the iLIDS-VID dataset.

Fig. 7 illustrates the CMC curves of different body parts. We observe that the performance of the torso part is greater than limbs. The reason is that the torso part is more discriminative for person ReID on the iLIDS-VID dataset. In addition, we improve performance by combining different body parts with attention scores, since these six body parts have structured complementary information.

D. Parameters Analysis

In this section, we conducted experimental analysis on the iLIDS-VID dataset to investigate the effect of parameter settings on proposed STAL: the sequence length, the rate in the fusion of global feature and local feature, margins m_l, m_g in Eq 4 and the feature embedding size.

1) Analysis of Fusion Rate: In our STAL model, we extract both global features and local body-part features to represent person video. When testing, we calculate the distance matrices of global and local features independently and fuse them with a balance rate as the final similarity metric. The rate is set empirically to improve the performance. Therefore, we investigated a few different fusion rates, which range from 0 to 1 with an interval of 0.1. The result performances on the iLIDS-VID dataset with different fusion rates are illustrated in Fig. 8. Obviously, the fusion strategy improves the performance significantly, since the global semantic representation and local body-part descriptor are complementary. Our model achieves the optimal rank 1 accuracy when the fusion rate is set to 0.3.

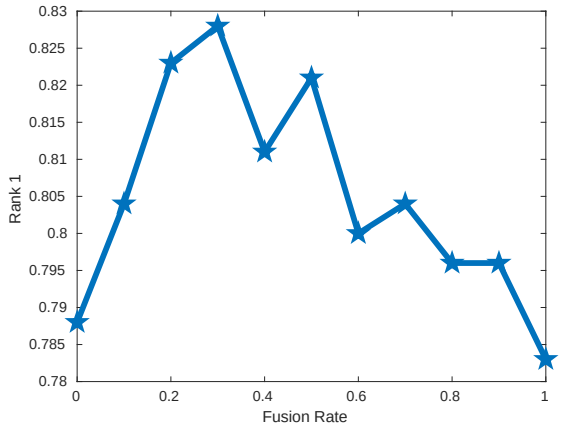


Fig. 8. Rank 1 accuracy on iLIDS-VID dataset with different fusion rates.

2) Analysis of Embedding Size: The embedding size of feature representation model is also crucial to the ReID problem. As shown in Fig. 9, we evaluated effects of different embedding sizes:{128,256,512,1024,2048} on the iLIDS-VID dataset. For the convenience of analysis, we independently investigated the global embedding size and local embedding size. For example, we fixed the local embedding size as 1024 when evaluating the effects of global embedding sizes, and vice versa. The performance constantly increases as embedding size increasing to 1024. While the growth is stagnant when the embedding size increases from 1024 to 2048. The reason may be that feature parameters are saturated when embedding size is in excess of 1024. The effects of both global embedding size and local embedding size are roughly consistency. The global embedding size is slightly sensitive than the local one. Finally, we choose the 1024 embedding size of both global and local feature in our experiments.

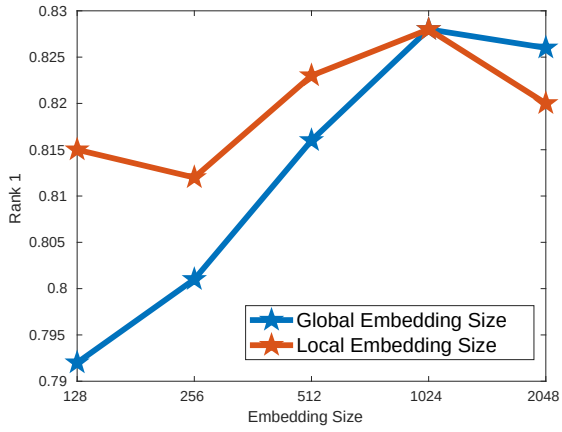


Fig. 9. Rank 1 accuracy on iLIDS-VID dataset with different embedding sizes.

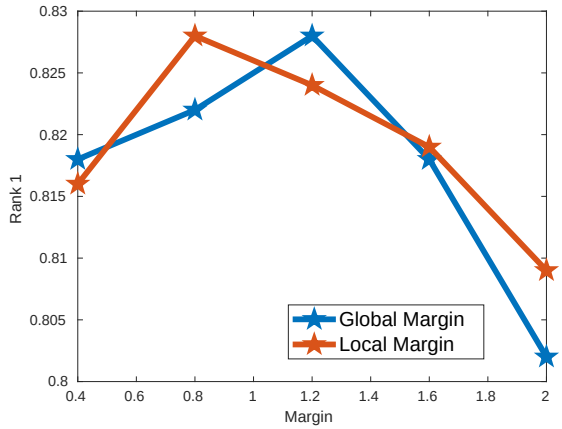


Fig. 10. Rank 1 accuracy on iLIDS-VID dataset with different margins in triplet loss.

3) Analysis of Margin: We trained proposed STAL with triplet loss, softmax loss, and consistency constraint as Eq. 3. The margins m_l, m_g in the triplet loss are crucial hyper-parameters which affect the generalization of the model by regularization. In this experiment, we investigated how re-identification accuracy varies depending on the margins of global distance and local distance. Evaluations were performed on the iLIDS-VID dataset with the margins varied from 0.4 to 2 with a 0.4 interval. When testing the global margin, the local margin was fixed at 0.8; while the global margin was set to 1.2 for evaluating the local margin. Fig. 10 illustrates the rank 1 accuracies on the iLIDS-VID dataset with different margins. We observe that the performance first increases corresponding to the margin and reaches the optimal around 1. Then, with the growth of margin, the performance drops gradually.

4) Analysis of Sequence Length: In this subsection, we investigated how the performance of the proposed method changes when the lengths of both probe and gallery sequences are varied. We trained the model on iLIDS-VID dataset as the original setting which selects 8 frames randomly in a video. During testing, we changed the lengths of both probe and gallery sequences from 1 to 128 in steps corresponding

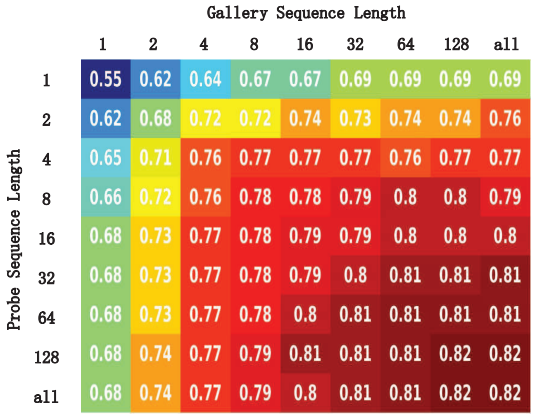


Fig. 11. Rank 1 CMC performance on iLIDS-VID dataset with different lengths of both probe and gallery sequences.

TABLE VI
RANK CMC ACCURACY OF CROSS-DATASET TESTING

Datasets	iLIDS-VID/PRID-2011			
Rank@R	R=1	R=5	R=10	R=20
CNN+RNN [15]	28.0	57.0	69.0	81.0
QAN [48]	34.0	61.3	74.0	83.1
ASTPN [62]	30.0	58.0	71.0	85.0
RQEN [63]	61.8	82.6	90.4	96.1
STAL	**63.7**	**84.0**	**92.8**	**98.1**

with the powers-of-two and give the reference which uses all frames. For instance, we fixed the sequence length in L. If the real length of a video is greater than L, we randomly selected L frames as the testing sequence; otherwise, we used the whole sequence and randomly sampled other frames to complement the L length. Different from the RNN-based method, we selected random frames instead of the first or last L frames of the consecutive sequence, since frames of a video with different poses and background have independent assuming in our model.

Results are reported in Fig 11 as a heat-map which shows the Rank 1 CMC accuracies with varied probe and gallery sequence lengths. It is easy to observe that the re-identification accuracy improves with the increase of the input sequence lengths of both probe and gallery videos. The detailed relations about accuracies and sequence lengths are divided into three stages: 1) the improvement of increasing frames is dramatic when the lengths of sequence $L \leq 4$; 2) the improvement is slight with the range of sequence lengths in $4 < L \leq 64$; and 3) the performance tends to be stable when the sequence lengths $L > 64$. It is understandable since the effectiveness of out temporal attention model is limited with few selected frames. It is difficult to select clear and informative frames and obtain robust representation when all frames are noise ones. In contrast, when the sampled frames have enough discriminative information, our STAL method aggregates the robust and discriminative representation with the attention-aware salience learning.

E. Robustness Analysis

In this section, we evaluated the robustness and generalization of our STAL method including cross-dataset evaluation and robustness evaluation about occlusions.

1) Robustness of Cross-Dataset Evaluation : In real surveillance systems, time and monetary cost are prohibitive to label overwhelming amount of data. To guide the training of model, we do need the labeled training data of the existing dataset. However, most of the existing experiments split the dataset into training and testing sets to evaluate the method. It may not be applicable in the real-world applications if the model learned by training data is over-fitting on the test data.

To evaluate the effects of the proposed method applied to a real-world surveillance system, we conducted the cross-dataset testing [15], where the model is trained by the iLIDS-VID dataset and tested on the PRID-2011 dataset. We also repeated the experiments 10 times and calculate the average accuracy.

Table VI summarizes the performances of the state-of-the-art methods in the cross-dataset testing, including CN-

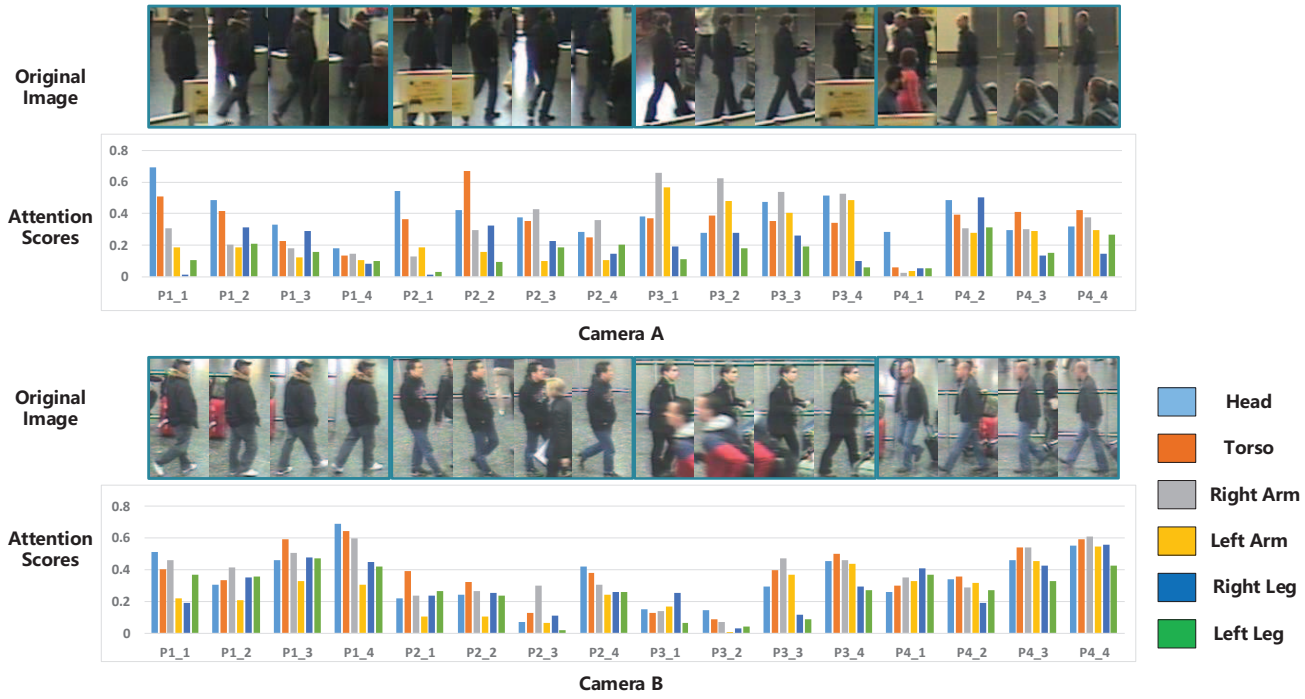


Fig. 12. Spatial-temporal attention scores of person samples in iLIDS-VID dataset. The top two rows and the bottom two rows show the attention scores of the same person video in two cameras, respectively. For a pedestrian video, we segment the person into six body parts and select four frames temporally.

Fig. 13. Samples of polluted iLIDS-VID datasets with random occlusions. The images of original dataset are occluded by a random 25×25 black box which simulates the natural occlusions.

N+RNN [15], QAN [48], ASTPN [62] and RQEN [63]. Compared with the single dataset testing, the performances of all methods on cross-dataset testing drop substantially. Our 29.7% improvement on Rank 1 than QAN [48] demonstrates the superior generalization ability of the proposed method. Compared with CNN+RNN [15] and ASTPN [62], proposed approach obtains a dramatical improvement with the more robust person representation and spatial-temporal attention model. Compared with RQEN, 2% improvement indicates the relation between temporal and spatial domains has great generalization ability.

2) Robustness about Occlusions: In the surveillance systems of the real world, pedestrians may be fully occluded, especially in the crowded environment. To intuitively evaluate the robustness of the proposed method in case of occlusions and background noise, we manually designed an experiment with a "polluted" dataset by a random black box. As shown in Fig. 13, we applied a 25×25 pixels black box in each frame of iLIDS-VID datasets to simulate the occlusions and background noise. When the black box is on the pedestrian, it brings in occlusion distractions. It also changes the image background when the black box is in the background. Both training and testing sets are processed with simulated occlusions in the experiments.

We conducted the occlusion robustness evaluation on the STAL and other baseline methods, QAN [48], RQEN [63] and the baseline CNN method. As shown in Table VII, our method outperforms the baseline method by 8% on Rank-1 accuracy, which shows the robustness of proposed method to occlusions and background noise. Our STAL method obtains a competitive accuracy on the "polluted" dataset, since the spatial attention module effectively spots the occluded parts and assigns them low attention scores to diminish the negative impact for the person matching.

F. Visualization of The Attention Model

In this section, we provided a visualization of how attention scores vary in both spatial and temporal domains corresponding to influences of the environment. As shown in Fig. 12, we depicted the spatial-temporal attention scores generated by

TABLE VII
RANK CMC ACCURACY OF ROBUSTNESS EVALUATION ABOUT OCCLUSIONS

Datasets	**Polluted iLIDS-VID**			
Rank@R	**R=1**	**R=5**	**R=10**	**R=20**
Baseline	70.2	89.9	93.5	97.9
RQEN [63]	70.7	87.7	92.6	96.2
QAN [48]	61.7	80.2	90.5	95.5
ASTPN [62]	55.2	78.7	89.8	94.1
STAL	**78.4**	**96.2**	**98.7**	**99.6**

the proposed STAL method of 4 testing persons in the iLIDS-VID dataset. For each person video, we sampled 4 frames in both cameras and calculated 6 spatial attention scores for each frame. We observe that the attention scores reliably reflect the qualities of different body parts in the frames. As shown in the first frame of person 2 in camera A, attention scores decrease when occlusions occur in corresponding parts. The scores of person 1 and person 3 in camera B respectively illustrate that attention scores are negatively correlated with the area of background noise and occlusions. Considering the temporal attention scores only, we observe that the less cluttered frame have higher scores, for example, the last frames of person 2 and person 3 in camera B get the highest scores.

V. Conclusions

In this work, we have proposed a STAL method which learns the attention on spatial and temporal dimensions to select informative frames and seek the salient parts to attend in person ReID. We develop an end-to-end network to integrate the global feature on person body and local spatial-temporal feature on discriminative body parts. We evaluated our method on three public video person re-identification datasets and demonstrated consistent improvement of our proposed approach over state-of-the-art methods. Our proposed method assumes that frames in a video are independent, rather than an ordered frame sequence, which neglects the temporal context like gait or motion. In the future, we will try to learn the context-aware temporal attention, which is more robust in the pedestrian video.

References

[1] 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*, pp. 1335–1344, 2016.

[2] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *CVPR*, pp. 2197–2206, 2015.

[3] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *CVPR*, pp. 3908–3916, 2015.

[4] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in *CVPR*, pp. 1239–1248, 2016.

[5] R. Zhao, W. Oyang, and X. Wang, "Person re-identification by saliency learning," *TPAMI*, vol. 39, no. 2, pp. 356–370, 2017.

[6] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *CVPR*, pp. 1249–1258, 2016.

[7] 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.

[8] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person re-identification by video ranking," in *ECCV*, pp. 688–703, 2014.

[9] S. Karanam, Y. Li, and R. J. Radke, "Sparse re-id: Block sparsity for person re-identification," in *CVPR Workshops*, pp. 33–40, 2015.

[10] S. Karanam, Y. Li, and R. J. Radke, "Person re-identification with discriminatively trained viewpoint invariant dictionaries," in *ICCV*, pp. 4516–4524, 2015.

[11] K. Liu, B. Ma, W. Zhang, and R. Huang, "A spatio-temporal appearance representation for viceo-based pedestrian re-identification," in *ICCV*, pp. 3810–3818, 2015.

[12] 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*, pp. 868–884, 2016.

[13] X. Zhu, X.-Y. Jing, F. Wu, and H. Feng, "Video-based person re-identification by simultaneously learning intra-video and inter-video distance metrics," in *IJCAI*, pp. 3552–3559, 2016.

[14] J. You, A. Wu, X. Li, and W.-S. Zheng, "Top-push video-based person re-identification," in *CVPR*, pp. 1345–1353, June 2016.

[15] N. McLaughlin, J. Martinez del Rincon, and P. Miller, "Recurrent convolutional network for video-based person re-identification," in *CVPR*, pp. 1325–1334, June 2016.

[16] J. Chen, Y. Wang, and Y. Y. Tang, "Person re-identification by exploiting spatio-temporal cues and multi-view metric learning," *IEEE SPL*, vol. 23, no. 9, pp. 998–1002, 2015.

[17] C. Gao, J. Wang, L. Liu, J.-G. Yu, and N. Sang, "Temporally aligned pooling representation for video-based person re-identification," in *ICIP*, pp. 4284–4288, 2016.

[18] M. Gou, X. Zhang, A. Rates-Borras, S. Asghari-Esfeden, M. Sznaier, and O. Camps, "Person re-identification in appearance impaired scenarios," in *BMVC*, pp. 1–10.

[19] 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 *CVPR*, July 2017.

[20] N. McLaughlin, J. M. del Rincon, and P. Miller, "Video person re-identification for wide area tracking based on recurrent neural networks," *TCSVT*, 2017.

[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.

[22] W. Zhang, X. Yu, and X. He, "Learning bidirectional temporal cues for video-based person re-identification," *TCSVT*, 2017.

[23] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," in *SCIA*, pp. 91–102, 2011.

[24] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato, "Hierarchical gaussian descriptor for person re-identification," in *CVPR*, pp. 1363–1372, 2016.

[25] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan, "End-to-end comparative attention networks for person re-identification," *TIP*, 2017.

[26] J. Lin, L. Ren, J. Lu, J. Feng, and J. Zhou, "Consistent-aware deep learning for person re-identification in a camera network," in *CVPR*, 2017.

[27] D. Li, X. Chen, Z. Zhang, and K. Huang, "Learning deep context-aware features over body and latent parts for person re-identification," in *CVPR*, July 2017.

[28] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang, "Spindle net: Person re-identification with human body region guided feature decomposition and fusion," in *CVPR*, 2017.

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

[30] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: A deep quadruplet network for person re-identification," in *CVPR*, 2017.

[31] Y. Duan, J. Lu, J. Feng, and J. Zhou, "Context-aware local binary feature learning for face recognition," *TPAMI*, vol. 40, no. 5, pp. 1139–1153, 2018.

[32] R. Zhao, W. Ouyang, and X. Wang, "Person re-identification by salience matching," in *ICCV*, pp. 2528–2535, 2013.

[33] J. Lu, V. E. Liong, and J. Zhou, "Deep hashing for scalable image search," *TIP*, vol. 26, no. 5, pp. 2352–2367, 2017.

[34] R. Zhao, W. Ouyang, and X. Wang, "Learning mid-level filters for person re-identification," in *CVPR*, pp. 144–151, 2014.

[35] J. Lu, V. E. Liong, and J. Zhou, "Simultaneous local binary feature learning and encoding for homogeneous and heterogeneous face recognition," *TPAMI*, vol. 40, no. 8, pp. 1979–1993, 2018.

[36] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *CVPR*, pp. 2288–2295, 2012.

[37] H. Trevor, T. Robert, and F. JH, "The elements of statistical learning: data mining, inference, and prediction," 2009.

[38] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian, "Local fisher discriminant analysis for pedestrian re-identification," in *CVPR*, pp. 3318–3325, 2013.

[39] J. Lu, J. Hu, and Y.-P. Tan, "Discriminative deep metric learning for face and kinship verification," *TIP*, vol. 26, no. 9, pp. 4269–4282, 2017.

[40] B. J. Prosser, W.-S. Zheng, S. Gong, T. Xiang, and Q. Mary, "Person re-identification by support vector ranking.," in *BMVC*, 2010.

[41] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *CVPR*, pp. 152–159, 2014.

[42] A. Subramaniam, M. Chatterjee, and A. Mittal, "Deep neural networks with inexact matching for person re-identification," in *NIPS*, pp. 2667–2675, 2016.

[43] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv*, 2017.

[44] W. Li, X. Zhu, and S. Gong, "Person re-identification by deep joint learning of multi-loss classification," *arXiv preprint arXiv:1705.04724*, 2017.

[45] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Deep attributes driven multi-camera person re-identification," in *European conference on computer vision*, pp. 475–491, 2016.

[46] A. Klaser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," in *BMVC*, pp. 1–10, 2008.

[47] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person re-identification by discriminative selection in video ranking," *TPAMI*, vol. 38, no. 12, pp. 2501–2514, 2016.

[48] Y. Liu, J. Yan, and W. Ouyang, "Quality aware network for set to set recognition," in *CVPR*, July 2017.

[49] Y. Yan, B. Ni, Z. Song, C. Ma, Y. Yan, and X. Yang, "Person re-identification via recurrent feature aggregation," in *ECCV*, pp. 701–716, 2016.

[50] V. Mnih, N. Heess, A. Graves, *et al.*, "Recurrent models of visual attention," in *NIPS*, pp. 2204–2212, 2014.

[51] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "From facial parts responses to face detection: A deep learning approach," in *ICCV*, pp. 3676–3684, 2015.

[52] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *ICML*, pp. 2048–2057, 2015.

[53] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang, "The application of two-level attention models in deep convolutional neural network for fine-grained image classification," in *CVPR*, pp. 842–850, 2015.

[54] M. Pedersoli, T. Lucas, C. Schmid, and J. Verbeek, "Areas of attention for image captioning," in *ICCV*, 2017.

[55] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *CVPR*, 2014.

[56] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *CVPR*, 2017.

[57] 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*, pp. 675–678, 2014.

[58] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *ICCV*, pp. 1116–1124, 2015.

[59] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *TPAMI*, vol. 32, no. 9, pp. 1627–1645, 2010.

[60] A. Dehghan, S. Modiri Assari, and M. Shah, "Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking," in *CVPR*, pp. 4091–4099, 2015.

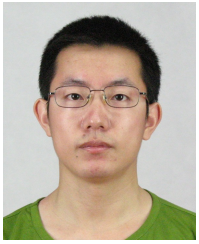
[61] G. Chen, J. Lu, J. Feng, and J. Zhou, "Localized multi-kernel discriminative canonical correlation analysis for video-based person re-identification," in *ICIP*, pp. 111–115, 2017.

[62] S. Xu, Y. Cheng, K. Gu, Y. Yang, S. Chang, and P. Zhou, "Jointly attentive spatial-temporal pooling networks for video-based person re-identification," in *ICCV*, 2017.

[63] G. Song, B. Leng, Y. Liu, C. Hetang, and S. Cai, "Region-based quality estimation network for large-scale person re-identification," *arXiv preprint arXiv:1711.08766*, 2017.

[64] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *CVPR*, 2017.

[65] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," *arXiv preprint arXiv:1708.04896*, 2017.



Guangyi Chen received the B.S. degree in the department of automation in Tsinghua University, China, in 2016. He is currently pursuing the Ph.D. degree at the Department of Automation, Tsinghua University. His research interests include person re-identification, video analysis, metric learning and deep learning.

Jiwen Lu (M'11-SM'15) received the B.Eng. degree in mechanical engineering and the M.Eng. degree in electrical engineering from the Xi'an University of Technology, Xi'an, China, in 2003 and 2006, respectively, and the Ph.D. degree in electrical engineering from Nanyang Technological University, Singapore, in 2012. He is currently an Associate Professor with the Department of Automation, Tsinghua University, Beijing, China. His current research interests include computer vision, pattern recognition, and machine learning. He has authored/co-authored over 200 scientific papers in these areas, where 60+ of them are IEEE Transactions papers (including 13 T-PAMI papers) and 50 of them are CVPR/ICCV/ECCV/NIPS papers. He serves the Co-Editor-of-Chief of the Pattern Recognition Letters, an Associate Editor of the IEEE Transactions on Image Processing, the IEEE Transactions on Circuits and Systems for Video Technology, the IEEE Transactions on Biometrics, Behavior, and Identity Science, and Pattern Recognition. He is a member of the Multimedia Signal Processing Technical Committee and the Information Forensics and Security Technical Committee of the IEEE Signal Processing Society, and a member of the Multimedia Systems and Applications Technical Committee and the Visual Signal Processing and Communications Technical Committee of the IEEE Circuits and Systems Society. He was a recipient of the National 1000 Young Talents Program of China in 2015, and the National Science Fund of China for Excellent Young Scholars in 2018, respectively. He is a senior member of the IEEE.

Ming Yang received the BE and ME degrees in electronic engineering from Tsinghua University, Beijing, China, in 2001 and 2004, respectively, and the PhD degree in electrical and computer engineering from Northwestern University, Evanston, Illinois, in June 2008. From 2004 to 2008, he was a research assistant in the computer vision group of Northwestern University. After his graduation, he joined NEC Laboratories America, Cupertino, California, where he was a senior researcher. He was a research scientist in AI Research at Facebook (FAIR) from 2013 to 2015. Now he is the co-founder and VP of software at Horizon Robotics, Inc. His research interests include computer vision, machine learning, face recognition, large scale image retrieval, and intelligent multimedia content analysis. He is the author of more than 50 peer reviewed publications in prestigious international journals and conferences, which have been cited over 9,000 times. He is a member of the IEEE.

Jie Zhou (M'01-SM'04) received the BS and MS degrees both from the Department of Mathematics, Nankai University, Tianjin, China, in 1990 and 1992, respectively, and the PhD degree from the Institute of Pattern Recognition and Artificial Intelligence, Huazhong University of Science and Technology (HUST), Wuhan, China, in 1995. From then to 1997, he served as a postdoctoral fellow in the Department of Automation, Tsinghua University, Beijing, China. Since 2003, he has been a full professor in the Department of Automation, Tsinghua University. His research interests include computer vision, pattern recognition, and image processing. In recent years, he has authored more than 100 papers in peer-reviewed journals and conferences. Among them, more than 30 papers have been published in top journals and conferences such as the IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Transactions on Image Processing, and CVPR. He is an associate editor for the IEEE Transactions on Pattern Analysis and Machine Intelligence and two other journals. He received the National Outstanding Youth Foundation of China Award. He is a senior member of the IEEE.