Learning Recurrent 3D Attention 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 propose to learn recurrent 3D attention (A3D) for video-based person re-identification. Attention model plays a key role in both spatial and temporal domains for video representation. Most existing methods apply spatial attention model to extract feature from a single image and aggregate image features with attentive temporal pooling or RNN. However, the inherent consistencies and correlations between spatial and temporal clues are not leveraged. Our A3D method aims to utilize the joint constraints of temporal and spatial attentions to enhance the robustness of attention model. Towards this goal, we treat the pedestrian video as a unified 3D bin where the temporal domain is denoted as an additional dimension. Then we develop an attention agent to iteratively select the locations of the salient spatial-temporal parts in the 3D bin. In addition, we formulate our sequential 3D attention learning as a Markov Decision Process and train the representation network and attention detector with the policy gradient method in an end-to-end manner. We evaluate the proposed method on three challenging datasets including iLIDS-VID, PRID-2011 and the large-scale MARS dataset, and consistently improve the performance in comparison with the state-of-the-art methods.**

***Index Terms*—Person re-identification, 3D attention, reinforcement learning, recurrent model.**

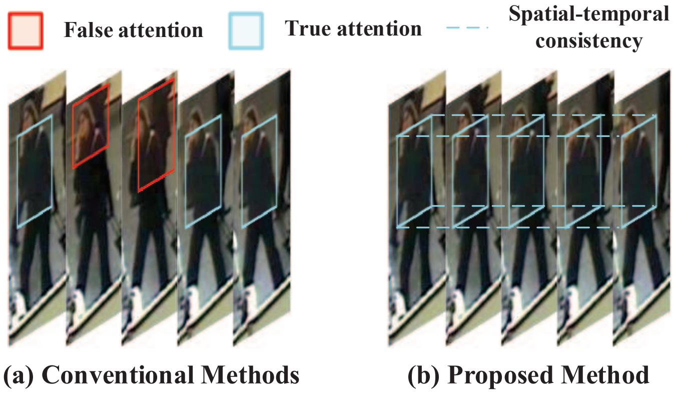


Fig. 1. Difference between conventional methods and our method. (a) shows that conventional attention methods explore the spatial attention region in each frame independently, which ignore the spatial-temporal consistency in the video. The learned attention regions of the consecutive frames are usually misaligned. (b) shows that the proposed A3D method jointly locates attentive 3D bins in a video clip, in which spatial-temporal consistency enhances the robustness of an attention detector. The temporal cues are applied to refine the misaligned attention regions as a constraint. (Best viewed in color)

I. Introduction

Person re-identification (ReID) aims to identify an individual across multiple non-overlapping camera views from a large set of candidates, with great potential in surveillance applications [1]. It is such an intriguing vision problem that the complicated inter-camera variances present all kinds of challenges, such as pose variations, illumination changes, partial occlusions, and clutter background.

In terms of the basic unit in the matching process, approaches for ReID are mainly divided into two categories: image-based methods [2]–[14] and video-based ones [15]–[24]. In this work, we investigate person re-identification with video sequences, which is a more practical setting than with still images, and gains increasing research interests recently [25]–[33]. In fact, surveillance videos with pedestrian are the original data for image-based person ReID before pre-processing. These person videos preserve abundant and potentially complementary spatial-temporal characteristics of pedestrian, from different poses and view angles. Recently, attention models have been successfully applied to video-based person ReID to explore salient regions [27], [34], select key frames [33], [35] and learn discriminative representations [26], [36]. Most existing video-based person ReID methods apply a spatial attention model to explore salient regions in a single image and employ temporal pooling with attention to aggregate features from multiple images. However, these methods capture spatial and temporal attentions separately, which ignore the inherent correlations between spatial and temporal clues.

Most pedestrian videos are located by person trackers or manual refinement, where temporal and spatial cues are interdependent. Spatial salient regions or occlusions move smoothly in the consecutive frames, which leads to the spatial-temporal consistency in the person video. While the person across frames may be misaligned, the locations of salient regions (*occlusions*) will not change dramatically due to the feasible moving speed of pedestrians and dense sampling of frames. For example, as shown in Fig. 1, the spatial

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, Grant U1813218, Grant U1713214, and Grant 61672306, in part by Beijing Academy of Artificial Intelligence (BAAI), in part by a grant from the Institute for Guo Qiang, Tsinghua University, and in part by the Shenzhen Fundamental Research Fund (Subject Arrangement) under Grant JCYJ20170412170602564. *(Corresponding author: Jiwen Lu).*

Guangyi Chen and Jiwen Lu are with the Beijing National Research Center for Information Science and Technology (BNRist), Department of Automation, Tsinghua University, Beijing, 100084, China. Email: chengy16@mails.tsinghua.edu.cn; lujiwen@tsinghua.edu.cn.

Ming Yang is with Horizon Robotics, Inc., Beijing & Shenzhen, 100080, China. E-mail: m-yang4@u.northwestern.edu.

Jie Zhou is with the Beijing National Research Center for Information Science and Technology (BNRist), Department of Automation, Tsinghua University, and the Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China. Email: jzhou@tsinghua.edu.cn.

salient regions of consecutive frames usually appear in the similar location. Learning the attention model for each frame individually in the video suffers from the position bias in the consecutive frames. Visual attention may be unreliable since the networks that generate them are often trained in a weakly-supervised manner. Therefore, the prior knowledge about spatial-temporal consistency in the video provides natural constraints for weakly-supervised attention learning.

Motivated by the observations that temporal and spatial attentions are inter-related, we propose a recurrent 3-dimensional attentive reinforcement learning framework, which treats the pedestrian video as a unified 3D bin and locates attentive 3D bins in it. In our A3D model, except for the key-frame selection and additional discriminative gait feature, the temporal information in the pedestrian video is applied to refine spatial attention as a constraint. With the 3D constraints, the temporal attention will focus on a consecutive video clip instead of discrete key frames. Specifically, we apply a two-stream network to extract static perceptual and dynamic motion information and obtain a global feature map. Then we propose a 3D attention detector, which selects salient bins in the feature maps and extracts local features. Inspired by the human vision system attending multiple salient objects sequentially, we employ the recurrent model in the attention generator to capture a sequence of salient 3D bins.

The sequential selection of 3D bins is a non-differentiable process, since the attention selection action lacks supervision signal. This motivates us to formulate spotting the attention 3D bins as a Markov Decision Process which is well optimized by reinforcement learning (RL) algorithms. In this formulation, the action is searching the center points of salient bins in the video, the state is the spatial-temporal features of the current bin and the hidden state from the previous iterations, and the reward measures the similarity rank and recognition accuracy by the final features of video clips. The RL algorithm optimizes each step of attention selection by setting explicit rewards towards person matching, rather than indirect weak supervision with classification loss, leading to a principled sequential attention learning. Besides, The RNN model is applied to replace the single attention bin with multiple ones for more robust representation learning. Similar with the process human vision system attending multiple salient objects sequentially, each step of RNN model indicts a glimpse of the human vision system. Experimental results on three video datasets including PRID-2011 [37], iLIDS-VID [15] and MARS [19] 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 recurrent 3-dimensional attentive (A3D) reinforcement learning framework to jointly attend to the salient parts of pedestrian videos on both temporal and spatial domains. We treat the pedestrian video as a unified 3D bin and seek a sequence of salient local bins, where the temporal attention and spatial attention are constrained in each local 3D bin.

2) We formulate the sequential selection of 3D video bins as a Markov Decision Process and design a reinforcement learning (RL) algorithm to end-to-end optimize our A3D attention model.

3) We conduct extensive experiments on three challenging video datasets including PRID-2011 [37], iLIDS-VID [15] and MARS [19] to demonstrate the efficacy of our A3D method. The results show that the proposed method outperforms other state-of-the-art methods.

II. Related Work

Existing ReID work cab be roughly divided into two categories: *image-based person ReID* and *video-based person ReID*. In this section, we first briefly review both image-based and video-based person ReID. Then, we will introduce two types of works: *attention model* and *reinforcement learning*, which are related to our approach.

A. Image-based Person Re-identification

Image-based person ReID systems [5], [7], [8], [25], [38]–[45] roughly consist of two components: robust representation learning and discriminative metric learning. Conventional representation learning methods [5], [39], [46], [47] try to extract a robust feature which is invariant to environmental and viewpoint changes. For example, LOMO [5] and GOG [39] are hand-crafted descriptors that combine the color and texture features. Salience match [46] learns image representation by seeking pairwise salient patch. Besides robust representation learning, metric learning [5], [6], [8], [48]–[52] also has been widely applied for person ReID. Previous methods [5], [49], [53], [54] learn a discriminant subspace or an integrated metric to emphasize inter-person distance and deemphasize intra-person distance. To learn the nonlinear relation of persons, the kernel-based metric learning methods [6], [55] are proposed, which project the feature vector from low dimension space to high dimension Hilbert space. Recently, deep neural networks have been applied to person ReID successfully to jointly learn feature representation and similarity within one network. In terms of loss functions, the deep learning based person ReID methods are categorized into three areas: verification, classification, and retrieval. Verification based works [9], [10], [56] apply the Siamese network to extract deep features and use the binary verification accuracy as supervisory signal. Classification based methods formulate person ReID as a classification problem and learn the deep discriminative features with the Softmax loss. For retrieval, a typical work is triplet loss [1], [12], [40], [45], which learns robust embeddings by preserving the rank relationship with a margin among each triplet of person samples.

B. Video-based Person Re-identification

Video sequences provide abundant person samples and their possible correspondences in consecutive frames. Thus, video-based person ReID methods [15]–[26] devote great efforts to extract robust spatio-temporal features to leverage the motion clues or aggregate image-based features along the time axis. To effectively leverage motion clues in the spatio-temporal feature learning, many previous methods [16], [17], [21], [57] employ the HOG3D [58] method on person videos as spatio-temporal features; Liu *et al.* [18] segment a video into a series of spatial-temporal body-action units and apply fisher vector to concatenate them in the final representation; and Chen *et al.* [28] and Liu *et al.* [30] employ optical flow to extract the motion information and capture dynamic gist. Then, to aggregate image-based features temporally, RFANet [59] employs a long short term memory (LSTM) network to capture contextual dependencies between the frames; while some recent works [22], [26] apply a temporal pooling layer following the recurrent model (RNN or LSTM) to capture long-term correlation in the sequence; Liu *et al.* [35], Xu *et al.* [36], and Li *et al.* [27] further improve the original temporal pooling by an attention model to select key frames adaptively.

C. Attention Model

Attention model [60] naturally imitates human perception to concentrate on what we are interested in. Recently, it gains great success in various fields, such as natural language processing (NLP), image understanding, and video analysis. Attention model also attracts increasing research interests in person ReID both in the spatial and temporal domains. In spatial domain, some works [40], [61]–[63] attempt to locate the informative spatial regions by generating attentive masks on the person image. Attention model is also employed to identify informative samples along the time axis. For example, QAN [35] designs a quality-aware network to estimate the image quality scores and integrates frame-wise features weighted by these scores. In addition, some works jointly learn the spatial and temporal attentions for robust feature learning. Xu *et al.* [36] design a pooling layer to jointly select spatial regions and temporal informative frames. Li *et al.* [27] weigh the features from different spatial regions and temporal frames by learning spatial-temporal quality scores. Different from these attention methods aggregating features with attentive weights, we directly detect 3D salient bins in the video, which adequately satisfy the constraints between spatial and temporal attentions.

D. Reinforcement Learning

Reinforcement learning (RL) aims to guide an agent to make optimal decisions by interacting with dynamic environment, which has been successfully applied in the vision tasks: object detection, visual tracking, and video analysis. Recently, RL has been adopted for person re-identification to generate spatial or temporal attention. For example, Zhang *et al.* [64] develop an agent to decide whether extra image pairs are needed, which achieves a reasonable trade-off between speed and accuracy. In contrast, SPL [65] employ deep RL to discard confounding frames from video. Xu *et al.* [34] learn a series of deformation actions of the bounding box to select the attention region of each person image. Some related works [60], [66] consider the spatial attention of image as the sequential decision process of an attentive agent interacting with a visual environment. In this paper, we extend this work in the video to locate 3D salient bins, which jointly learns both spatial and temporal attentions. Meanwhile, we address the problems about a huge action space, high computational cost and temporal importance bias in the video.

III. Approach

In this section, we propose a recurrent 3D attentive (A3D) reinforcement learning framework that treats the pedestrian video as a unified 3D bin and develop an attention agent to iteratively select locations of the salient spatial-temporal parts. We first present the overall architecture of the proposed method and then explain our recurrent 3D attention model in details. Finally, we explain the optimization procedure and implementation details of our A3D method.

A. Overall Procedure

Fig. 2 illustrates the overall procedure of the proposed A3D method. Inspired by [67], we also propose a two-steam representation network for video person ReID to better explore the dynamic motion in the video sequence. Given a pedestrian video $X \in R^{3\times L\times H\times W}$, where L is the number of video frames and $H \times W$ is the spatial size of each frame, the optical flows between adjacent video frames calculated by Flownet [68] are denoted as $F \in R^{2\times L-1\times H\times W}$. To ensure the video length consistent between the original RGB-based video clip and generated optical flow sequence, we repeat the last optical flow map as the temporal padding. Then, the original RGB-based video clips and generated optical flow maps are fed into the low-level ConvNet respectively to extract perceptual and motion features. Different from the RGB-based video, the first convolution layer of optical flow has two channels, which consist of the vertical and horizontal channels. While the dimension of output feature maps in the optical-flow stream are the same as the RGB-based ones. In addition, we merge the features from two streams by element-wise addition and feed them into the following high-level ConvNet as:

$$g = \mathcal{F}_g(\mathcal{F}_{img}(X) \otimes \mathcal{F}_{flow}(F)), \tag{1}$$

where $g \in \mathbb{R}^{C_g\times L_g\times H_g\times W_g}$ is the final global feature map.

Then we further employ an RNN-based 3D attention detector to obtain salient bins and local discriminative features iteratively. The input of our recurrent A3D attention model is the global features maps g, and the outputs are a sequence of local features extracted from different attention bins. It is expressed as:

$$\{f_l^t\}_{t=1:T} = \alpha(g) : \mathbb{R}^{L_g\times C_g\times H_g\times W_g} \mapsto \mathbb{R}^{T\times C_l}, \tag{2}$$

where, α denotes our A3D module, T is the number of glimpses in the attention model, and f_l^t is the local feature of tth attention glimpse. Meanwhile, we generate the global feature f_g using average pooling based on the global feature maps g. In the end, we concatenate the global feature f_g and local feature sequence $\{f_l^t\}_{t=1:T}$ to obtain the final video representation. To address the problem of limited computing resource and varying video length, we slice the video into multiple clips in both training and testing procedures. During training, we randomly select video clips to train our network. While in the testing procedure, we sequentially select clips of one video with a fixed stride, and average these features as the final video representation.

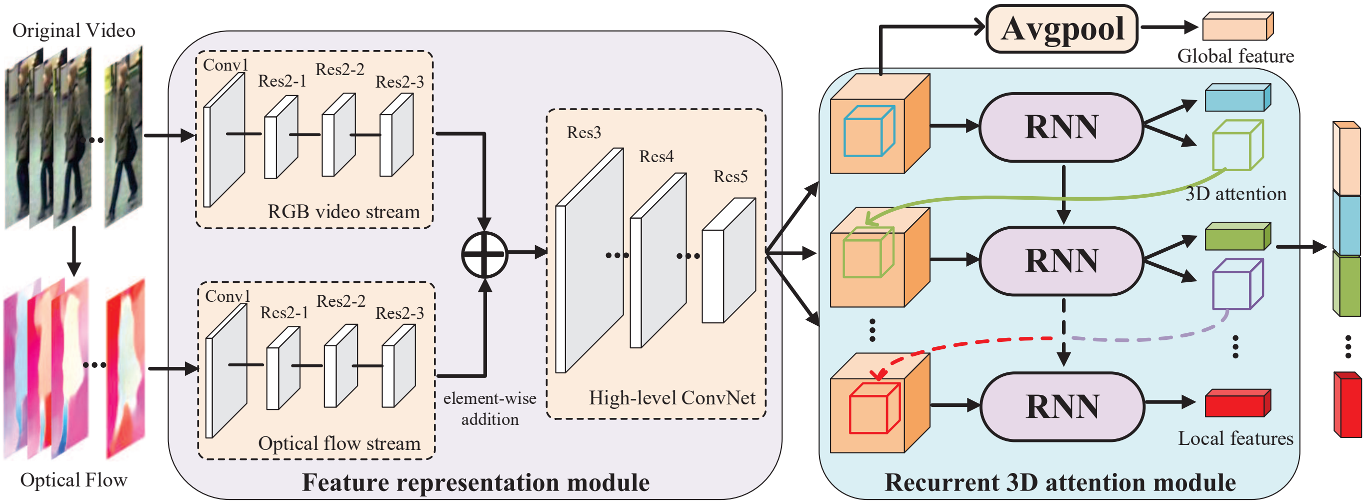


Fig. 2. The network architecture of the proposed recurrent A3D framework. The feature representation module includes two streams: RGB video stream for static perceptual features and optical flow stream for dynamic motion features. The two streams are merged by element-wise addition and fed in high-level ConvNet to extract global features. The recurrent 3D attention module iteratively selects salient bins from video and extract local features by an RNN network. The attention-based local features and global feature are connected as the final video representation. (Best viewed in color)

B. Recurrent 3D Attention Module

To jointly explore salient regions and key frames, we treat the pedestrian video as a 3D bin and develop a recurrent 3D attentive agent to iteratively locate the coordinates of informative spatial-temporal parts. This selection procedure of attentive 3D bins is non-differentiable in the spatio-temporal domain, which is hard to learn with back-propagation type of training. We therefore formulate our recurrent A3D model as a RL problem. The RL involves $\{\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}\}$ as the states, actions, transitions and rewards. At tth iteration, our A3D agent takes a state $s_t \in \mathcal{S}$ as the observation and predicts an action $a_t \in \mathcal{A}$, which indicates the location of salient bin. After that, the state is updated by the state transition distribution $\mathcal{T}(s_{t+1}|s_t, a_t)$. Until the maximum iteration is reached, the agent repeats the attention prediction and generates the local feature based on the selected salient bin. Finally, we integrate the global feature and local features based on all glimpse as the final feature of the video. Note that the relations between learned attention bins are complementary with the equal level but not sequentially refined. For example, the first attention focuses on the salient bag of the person, while the next one tends to focus another salient region, like the pants, rather iteratively refine the location of the salient bag. To train the agent to select appropriate actions, a reward $r_t \in \mathcal{R}$ is fed back from the environment, which evaluates selected attention bins with the generated features. We will elaborate the definitions of these states, actions, transitions, and rewards in the following.

States and Transitions: The state s_t of tth iteration includes two parts: the observation with current attention and the information extracted from previous iterations. With the current location generated by the last step, we extract the local feature of the selected attention bin. To reduce the computational cost of feature extraction, we crop our attention bin from high-level feature maps g, instead of the original input pedestrian video and optical flow. We share the parameters of most convolution layers when extracting the features of different attention bins, which significantly improves the efficiency during both training and test procedures. As shown in Figure 3,we formulate the observation as

$$o_t = \mathcal{F}_o(g, a_{t-1}), \tag{3}$$

where a_{t-1} represents the selected action which indicates attention location. On the other hand, the history of previous iterations also provides a sensible clue to guide the agent to make appropriate decisions. Therefore, we employ the hidden state of last iteratively h_{t-1} in our state s_t to "remember" the status of the past iterations. The supplement of history information allows the agent to analyze the contextual information among all selected attention bins. Finally, we formulate the internal state by a hidden unit of the RNN, which contains both the current observation and history information, as follows:

$$s_t = h_t = \mathcal{F}_h(o_t, h_{t-1}). \tag{4}$$

In this process, the internal state s_{t-1} transforms to s_t by the selected action a_{t-1}. As shown in Figure 3, the hidden embedding module $\mathcal{F}_h$ generates the current hidden embedding with the local observation o_t and last one h_{t-1}.

Actions: In our A3D method, the actions of the agent are all candidate attention bins over the high-level feature maps. The size of an action pool is a crucial hyper-parameter that influences the convergence and performance in the RL. Compared with the conventional image problem, pedestrian video always has a lager action space. To reduce the number of actions, instead of the original video data, we crop the salient bins in the high-level feature maps. In addition, we fix the scales of attention bins and only select the bins inside. For example, when we fix the size of the attention bin as $L_a \times H_a \times W_a$, and the global feature maps $g \in \mathbb{R}^{C_g \times L_g \times H_g \times W_g}$, we get the size of the action pool as $(L_g - L_a + 1)(H_g - H_a + 1)(W_g - W_a + 1)$. Given the state s_t, the action a_t of our A3D agent is denoted as $\{a_t = (l, h, w)|0 \le l \le L_g - L_a, 0 \le h \le H_g - H_a, 0 \le w \le W_g - W_a\}$. To encourage more exploration of location selection, at time step t, we stochastically select attention bins

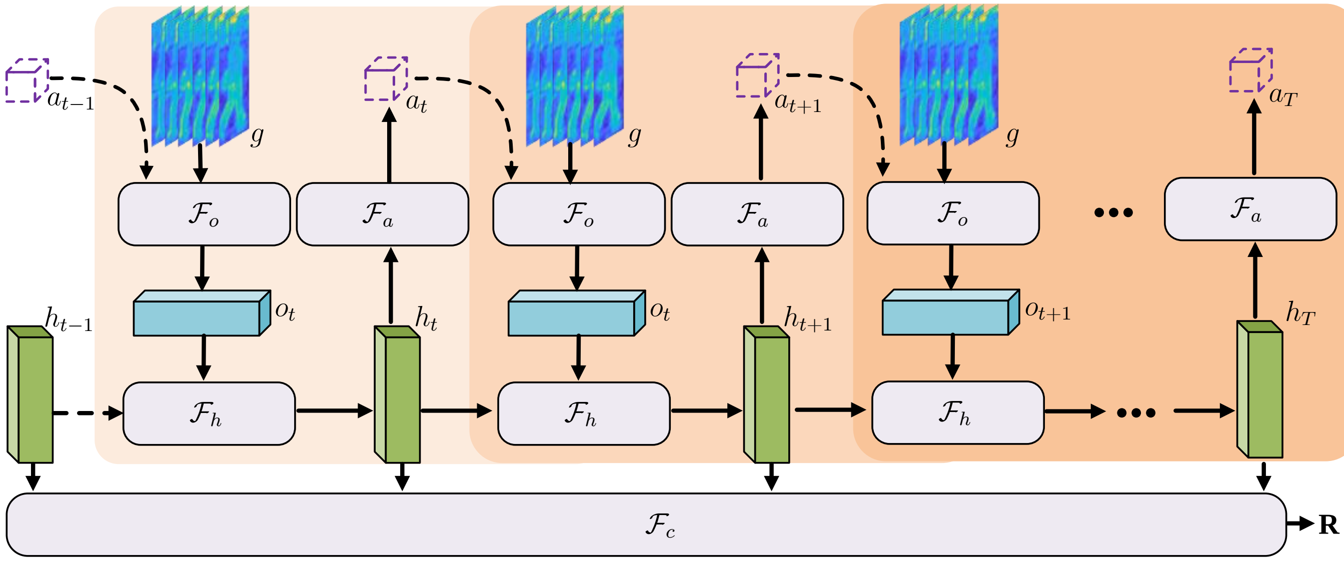


Fig. 3. The detail network architecture of the A3D attention learning process. The sequential attention learning is a Markov Decision Process formulated by an RNN model. The RNN model consists of four modules: observation module $\mathcal{F}_o$, hidden embedding module $\mathcal{F}_h$, attention locating module $\mathcal{F}_a$, and classification module $\mathcal{F}_c$. In tth step, the 3D attention predicted in last step a_{t-1} and the high-level representation g are fed into observation module $\mathcal{F}_o$ to get local observation o_t. Then $\mathcal{F}_h$ generates hidden embedding h_t with o_t and the last hidden embedding h_{t-1}. Attention locating module $\mathcal{F}_a$ predicts next 3D attention region with h_t. Finally, all hidden embedding are fed into the classification module $\mathcal{F}_c$ as local features for identifying person and obtaining rewards. (Best viewed in color)

over a Gaussian distribution which is parameterized by our attention prediction network:

$$a_t \sim \mathcal{N}(a_t|\mu = \mathcal{F}_a(h_t), \Sigma), \tag{5}$$

where $\mathcal{F}_a(\cdot)$ is the attention prediction network implemented with a fully-connected layer. For example, when the agent predicted a attention bin $\mathcal{F}_a(h_t)$ in the training stage, the policy explores a new attention bin from the neighborhood and determines if it's a better attention. While in the testing stage, we directly use the predicted result (the expectation) as the attention, which removes the exploration. The variance is a fixed parameter which denotes the strength of exploration. The lager variance indicts to encourage more exploration. We set the variance as $\Sigma = 0.03$ in all experiments. To avoid the mismatching with random attention initialization, we uniformity initialize the first glimpse as the center of the video clip and fill the initial hidden state with zeros.

Rewards: To guide the agent to make appropriate decisions, we design a reward signal r_t in each iteration, which reflects our task objective. For the objective of video-based person ReID, we take the recognition results as the reward of our A3D model to help the agent to select salient 3D bins which are beneficial to person matching. Formally, at tth iteration, we take the internal state $f_l^t = h_t$ as the local features based on last attention bins a_{t-1}. Meanwhile, we utilize an average pooling layer on the global feature maps to generate the global feature f_g. As shown in Figure 3, we employ a classification network to predict the identity of the given pedestrian sample, whose input is the connection of both global feature and local features:

$$p_c = \mathcal{F}_c(f_g, f_l^1, \ldots, f_l^T), \tag{6}$$

where p_c denotes the predicted probabilities of identity classes. Instead of only choosing the local feature of the last recurrent unit, we utilize the features of all iterations as local features to reduce sequential importance bias. Then we estimate the prediction label l_p of given pedestrian video clip by selecting the maximum of p_c. Suppose the ground-truth label is l_g, we define the reward at time step t as :

$$r_t = \begin{cases} 1 & l_g = l_p, t = T \\ 0 & otherwise \end{cases}. \tag{7}$$

In fact, things encouraging the agent to select key 3D bins can be employed as rewards. Then the goal of the agent is to maximize the sum of discounted reward as:

$$R = \sum_{t=1}^{T} \gamma^{t-1} r_t, \tag{8}$$

where γ is the discount factor, which is set as 1 in our experiments.

C. Optimization

During training, we learn the parameters of our A3D agent by maximizing the received reward in the person ReID. Following the REINFORCE [69], we define a policy as

$$P(a_t|s_t, \theta) \sim \pi_\theta(a_t|s_t), \tag{9}$$

where θ indicates the parameters of our agent. The policy π predicts a distribution over actions based on current state with parameters θ. Then, we denote τ as a trail which describes a sequence of past states and attentions, e.g. $\tau = \{s_1, a_1, \ldots, s_T, a_T\}$. We calculate the likelihood probability of the sequential decisions τ with the state transferring probability $\mathcal{T}$ and action selection probability π as:

$$P(\tau|\theta) = \prod_{t=1}^{T} \mathcal{T}(s_{t+1}|s_t, a_t)\pi_\theta(a_t|s_t). \tag{10}$$

Under this distribution of trails, we define the objective function with the maximization of expected reward as:

$$J(\theta) = \mathbb{E}_{P(\tau|\theta)} \sum_{t=1}^{T} r_t = \mathbb{E}_{P(\tau|\theta)} R. \tag{11}$$

To optimize the policy network π, we maximize the expected reward by repeatedly estimating the policy gradients with the Monte Carlo method as:

$$\begin{aligned} \nabla J(\theta) &= \mathbb{E}_{P(\tau|\theta)} \big[\nabla_\theta \log P(\tau|\theta) R\big] \\ &= \mathbb{E}_{P(\tau|\theta)} \big[\sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t) R\big] \\ &\approx \frac{1}{M} \sum_{m=1}^{M} \sum_{t=1}^{T} \big[\nabla_\theta \log \pi_\theta(a_t^m|s_t^m) R\big], \end{aligned} \tag{12}$$

where $m = 1, \ldots, M$ denotes the index of M episodes sampled in the Monte Carlo method. Although the above gradient estimator is simple and unbiased, it still suffers from the high variance problem. Therefore, we introduce a baseline to reduce the variance of our gradient estimation:

$$\nabla J(\theta) \approx \frac{1}{M} \sum_{m=1}^{M} \sum_{t=1}^{T} [\nabla_\theta \log \pi_\theta(a_t^m|s_t^m)(R - b)], \tag{13}$$

where b is a baseline which is independent of the action. By adding this baseline, we reduce the variance of gradient estimation without changing the expectation of the original gradient. Specifically, we employ the value function in the RL as our baseline, which is implemented by a value network.

In addition, to assist the training of A3D agent and learn discriminative video representation for person ReID, we add a hybrid supervised signal for the re-identification objective. The designed loss function consists of two parts: triplet loss and classification loss. The first triplet loss function aims to preserve the rank relationship among a triplet of pedestrian videos. It is formulated as:

$$L_{tri} = \frac{1}{N} \sum_{i=1}^{N} \big[||f_i - f_i^{+}||_2^2 - ||f_i - f_i^{-}||_2^2 + m\big]_+, \tag{14}$$

where $\big[x\big]_+$ denotes the max function $max(0, x)$, and f_i, f_i^+, f_i^- respectively denote as anchor sample, positive sample and negative sample in a triplet. m is a margin to enhance the discriminative ability of learned features, which is set to 0.3 in the experiments. The other classification loss concerns whether a given person is correctly identified. Specifically, we employ a cross entropy loss as:

$$L_{cls} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} y_i^c \log(p_i^c), \tag{15}$$

where y_i^c is the ground truth identify of ith person on the cth class and p_i^c indicates the predicted probability of our model.

Algorithm 1 : The A3D method

```
Input: The high-level feature maps $g$ of the sampled video clip, the size of the 3D attention region $L_a \times H_a \times W_a$, the glimpse number of the sequential attention model $T$.
Output: The parameters of observation module $\mathcal{F}_o$, hidden embedding module $\mathcal{F}_h$, attention locating module $\mathcal{F}_a$, and classification module $\mathcal{F}_c$.
Initialize the parameters of $\mathcal{F}_o$,$\mathcal{F}_a$,$\mathcal{F}_c$, and $\mathcal{F}_h$;
for $t = 0, 1, 2, \ldots, T$ do
    if $t = 0$ then
        Initial attention location $a_0$ as the center: $a_0 = (\lceil \frac{L_g - L_a}{2} \rceil, \lceil \frac{H_g - H_a}{2} \rceil, \lceil \frac{W_g - W_a}{2} \rceil)$
        Initial hidden state $h_0$ with zeros $h_0 = \mathbf{0}$
    else if $1 \leq t < T$ then
        Obtain local observation $o_t$ with the feature map $g$ and the attention region of the last glimpse $a_{t-1}$ as (3)
        Obtain hidden embedding $h_t$ with observation $o_t$ and last hidden state $h_{t-1}$ as (4)
        Predict attention location $a_t$ with hidden embedding $h_t$ as (5)
    else
        Classify the identity of the person video as (6)
        Obtain the rewards as (7) and (8)
        Calculate the reinforcement learning gradients as (13)
        Calculate the hybrid supervised gradients as $\nabla(L_{tri} + L_{cls})$
        Update the parameters of $\mathcal{F}_o$,$\mathcal{F}_a$,$\mathcal{F}_h$, and $\mathcal{F}_c$
    end if
end for
return The parameters of $\mathcal{F}_o$,$\mathcal{F}_a$,$\mathcal{F}_c$, and $\mathcal{F}_h$
```

TABLE I
DETAILED STRUCTURE OF RGB IMAGE BRANCH $\mathcal{F}_{img}$.

name	**kernel**	**stride**	**output size**
input			3 × 256 × 128
conv1	7 × 7	2	64 × 128 × 64
maxpool1	3 × 3	2	64 × 64 × 32
{res1_img}	1 × 1 3 × 3 × 3 1 × 1	1	256 × 64 × 32

D. Implementation Details

In our recurrent A3D reinforcement learning framework, the model includes 7 networks $\{\mathcal{F}_g, \mathcal{F}_{img}, \mathcal{F}_{flow}, \mathcal{F}_c, \mathcal{F}_o, \mathcal{F}_h \mathcal{F}_a\}$, where the first three networks aim to extract global appearance and motion features, and the others devote great efforts to seek salient 3D bins in the video. The backbone network architecture of feature representation model is ResNet50 [70]. We modify the basic network architecture to adapt to our video-based person ReID task. First, to extract the motion features of video, we add an independent optical flow stream which consists of 10 convolution layers. In order to merge

TABLE II
DETAILED STRUCTURE OF OPTICAL FLOW BRANCH $\mathcal{F}_{flow}$.

name	**kernel**	**stride**	**output size**
input			2 × 256 × 128
conv1	7 × 7	2	64 × 128 × 64
maxpool	3 × 3	2	64 × 64 × 32
{res1_flow}	1 × 1 3 × 3 × 3 1 × 1	1	256 × 64 × 32

TABLE III
DETAILED STRUCTURE OF THE HIGH-LEVEL CONVNET $\mathcal{F}_g$.

name	kernel	stride	output size
input			$256 \times 64 \times 32$
{res2}	1×1 3×3 $\times 6$ 1×1	2	$512 \times 32 \times 16$
{res3}	1×1 3×3 $\times 4$ 1×1	2	$1024 \times 16 \times 8$
{res4}	1×1 3×3 $\times 3$ 1×1	1	$2048 \times 16 \times 8$

TABLE IV
DETAILS OF THE RECURRENT 3D ATTENTION MODEL.

name	input size	output size
	$\mathcal{F}_o$	
input_feature		$2048 \times 8 \times 16 \times 8$
input_att		3×1
pool_ feature	$2048 \times 4 \times 6 \times 3$	2048×1
fc1_feature	2048×1	128×1
fc1_att	3×1	128×1
fc2_feature	128×1	256×1
fc2_att	128×1	256×1
addtion	$\{256 \times 1\} \times 2$	256×1
	$\mathcal{F}_h$	
fc_i2h	256×1	256×1
fc_h2h	256×1	256×1
	$\mathcal{F}_a$	
fc_att	256×1	3×1
	$\mathcal{F}_c$	
fc_cls	$2048 + 256 \times T$	number of classes

the optical flow stream network with the original RGB-based network by an element-wise addition, we apply the same network structures of two streams except for the input shape. Then, in order to locate the attention bins accurately from global feature maps, we expand the size of feature map by applying a convolution layer with $stride = 1$, instead of original $stride = 2$ convolution layer in ResNet50. $\mathcal{F}_c, \mathcal{F}_o, \mathcal{F}_h$ are implemented by fully-connected layers in the RNN model, whose hidden layers are 256 dimensions.

All CNN models in this work are pre-trained on ImageNet and fine-tuned with video person re-identification datasets. In the training stage, e.g., the iLIDS-VID and PRID-2011 datasets, we train our model on a single GTX 1080 Ti GPU machine for 500 epochs by Adam optimizer. The initial learning rate is 0.0002 and reduces by half with each 60 epochs. We randomly select 16 video clips from 4 persons in a batch, where each clip consists of 6 frames (for MARS, we train the model on 2 GPUs for 800 epochs with 48 clips from 12 persons in a batch). For the inputs of the original RGB-based video data and the optical-flow data, we apply randomly mirror and erase as data augmentation and resize them to 256×128. In the testing stage, we choose the Euclidean distance as the metric and calculate the similarities between probe and gallery samples. We sequentially extract the features of clips and average these features as the final video representation, where clips in the testing stage also include 6 frames and the stride is set to 0.

E. Network Architectures

TABLE V
THE BASIC STATISTICS 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

In this subsection, we introduce the structures of our backbone network and attention model in the detail. First, we describe the base model which extracts the global appearance and motion features with CNNs, including the image branch $\mathcal{F}_{img}$, optical flow branch $\mathcal{F}_{flow}$, and high-level ConvNet $\mathcal{F}_g$. As shown in Tables I-III, we list the kernel size, stride and output size of each layer. We employ two CNN branches to respectively capture the appearance and motion information and fuse them by element-wise addition. We feed the fused features into the high-level ConvNet $\mathcal{F}_g$ to extract the global features and apply an average pooling layer to aggregate the features from all frames as the video representation. The backbone network is based on the ResNet50 [70]. Different with original ResNet50, we add an independent branch to capture the gait clues and replace the original $stride = 2$ in the *res4* block with $stride = 1$ to expand the resolution of feature maps.

Second, we describe the details of our recurrent 3D attention model. Human vision system generally attends to multiple salient objects one by one. This inspires us to employ the RNN as the basic framework to *sequentially* capture salient local clues in a person video. Different from common usages of RNN only in the temporal domain, we adopt the RNN to model the sequential decision process of spotting 3D attention locations. We split the RNN cell into 4 modules including $\mathcal{F}_o$, $\mathcal{F}_h$, $\mathcal{F}_a$, and $\mathcal{F}_c$. We show the network structure in Figure 3 and illuminate the network details in Table IV. The inputs of the attention model consist of the global feature maps and attention coordinates. The $\mathcal{F}_o$ extracts the local attentive features and feeds them into $\mathcal{F}_h$. $\mathcal{F}_h$ takes the previous hidden state and current local features as input to learn the current state. $\mathcal{F}_a$ predicts the next attention location with a fully connected layer based on the current state. Finally, $\mathcal{F}_c$ uses both local and global features for person classification. T in Table IV denotes the number of steps in the recurrent model.

IV. EXPERIMENTS

We evaluated our method on three public pedestrian video datasets: iLIDS-VID [15], PRID-2011 [37], and MARS [19]. We compared the proposed method with other state-of-the-art approaches, and conducted ablation experiments and parameters analysis to investigate the effectiveness and robustness of the proposed A3D model.

A. Experimental Settings

Datasets: The detailed statistics of all datasets are presented in Table V. The iLIDS-VID dataset [15] contains 600 videos of 300 subjects, which have variable numbers of frames from

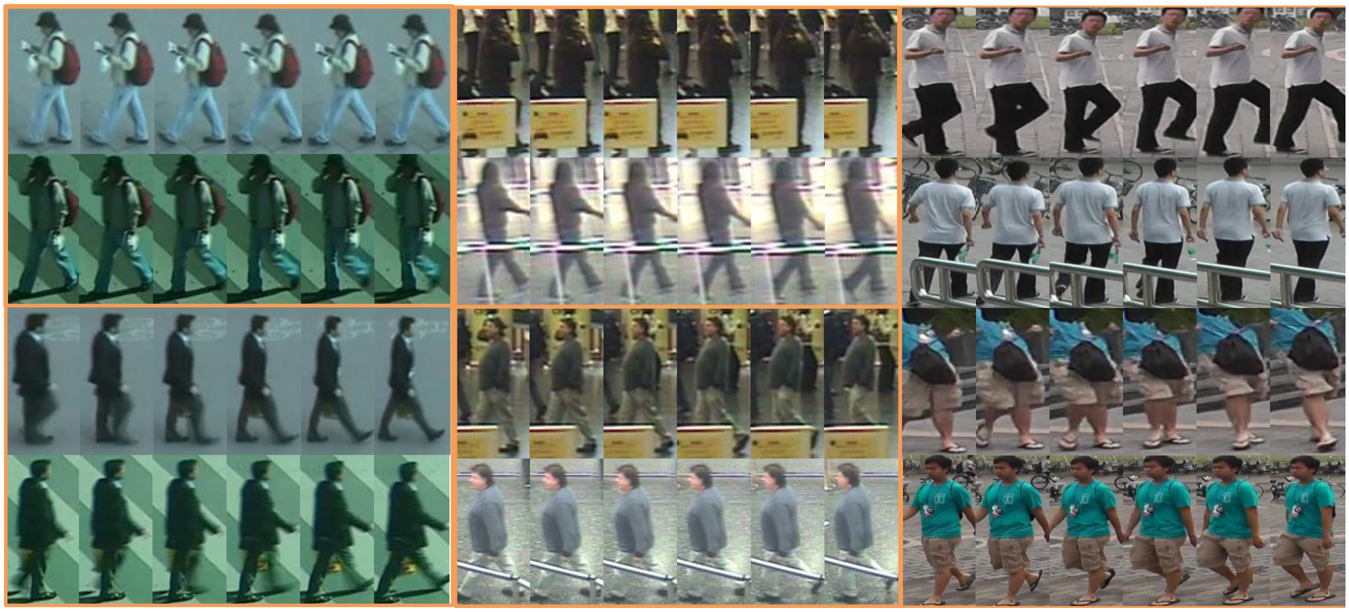


Fig. 4. Samples of iLIDS-VID, PRID-2011 and MARS datasets. The left part shows the video from PRID-2011, the middle one is sampled from iLIDS-VID, and the right part is the pedestrian samples from MARS. Each part consists of two individuals under different camera views, and we select 6 frames for each person video.

23 to 193 with an average of 73. As shown in Figure 4, the videos in iLIDS-VID dataset are captured in a crowded airport arrival hall with cluttered background and partial occlusion. The PRID-2011 dataset [37] consists of 385 persons in camera A and 749 persons in camera B. 400 videos of 200 pedestrians appear in both cameras. Following [15], we selected videos of 178 identities with more than 27 frames[1] in the experiment. As shown in Figure 4, the main challenge of the PRID-2011 dataset is illumination variance across two cameras views. MARS [19] is the largest video-based person ReID dataset with 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. The bounding boxes in MARS are generated by a DPM detector [71] and a GMMCP tracker [72], instead of the hand-drawn way. As shown in Figure 4, the incomplete frames and misalignment across different video clips due to the low-quality detector and tracker are the main challenges of the MARS dataset. Although the person videos are noisy and misaligned, we find that the locations of salient regions (*occlusions*) will not change much in the datasets, which motivates the proposed A3D model.

Experimental Setup: We followed the protocol of [15] for PRID-2011 and iLIDS-VID datasets. We repeated experiments 10 times and calculated the average accuracy by splitting the dataset into equal-sized training and testing sets. To avoid the noise from dataset splitting and make sure a fair evaluation, we selected the identical 10 splits in [15], instead of random splits. For MARS dataset, we followed the experimental setup in [19], which uses 625 persons for training and the others for testing. We applied cumulative matching characteristic (CMC) curve and mean Average Precision (mAP) as the evaluation metric. CMC curves record the true matching within the top n ranks, while mAP considers precision and recall to evaluate the overall performance of methods.

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

We compared the proposed approach with other state-of-the-art methods which include *conventional representation learning and metric learning methods*, such as AFDA [73], DVDL [17], mvRMLLC+ST+Alignment [23], STFV3D+KISSME [18], eSDC+MSSDALF+DVR [57], T-DL [21], AvgTAPR [24], LOMO+KISSME+SRID [74], LOMO+LMKDCCA [31]; *deep learning based methods* including CNN+RNN [75], CNN+XQDA [19], AMOC [30], FANet+RSVM [59], DeepRCN+KISSME [76], BRNN [77] and TRL [32]; and *attention-based methods* such as TAM+SRM [26], SDM [64], QAN [35], ASTPN [36], DSAN [29], SPL [65], DRSTA [27], CSSA+CASE [28], RQEN [78], STAL [33]. In addition, we also compare our A3D approach with two methods which also leverage reinforcement learning for video based person ReID problem, i.e. SDM [64] and SPL [65].

Table VI illustrates the performance of our A3D approach and most existing video-based person ReID approaches on iLIDS-VID, PRID-2011, and MARS datasets. The top group of Table VI shows results of methods which directly learn video representation without attention mechanism. While the bottom group shows the performance of deep learning methods with attention model on the temporal or spatial domain. We observe that the proposed A3D method consistently improves over all comparing methods substantially on the three benchmarks. For iLIDS-VID and PRID-2011 datasets, we improve the second best method by 2.5% and 2.1% respectively on the Rank-1 accuracy. While for large-scale MARS dataset, we obtain the comparable Rank-1 accuracy with CSSA+CASE [28], while achieve 3.7% improvement on the mAP metric.

[1] In PRID-2011, these 178 identities do have more than 27 frames, instead of reported 21 frames

TABLE VI
COMPARISON WITH THE STATE-OF-THE-ART PERSON RE-IDENTIFICATION METHODS ON THE iLIDS-VID, PRID-2011 AND MARS DATASETS.

Datasets	**iLIDS-VID**				**PRID-2011**				**MARS**		
Rank@R	**R=1**	**R=5**	**R=10**	**R=20**	**R=1**	**R=5**	**R=10**	**R=20**	**R=1**	**R=5**	**mAP**
AFDA [73]	37.5	62.7	73.0	81.8	43.0	72.7	84.6	91.9	-	-	-
DVDL [17]	25.9	48.2	57.3	68.9	40.6	69.7	77.8	85.6	-	-	-
mvRMLLC+ST+Alignment [23]	69.1	89.9	96.4	98.5	66.8	91.3	96.2	98.8	-	-	-
STFV3D+KISSME [18]	44.3	71.7	83.7	91.7	62.5	83.6	89.9	92.0	-	-	-
DynFV+LDFV [25]	28.8	55.0	70.6	82.0	43.6	69.0	79.4	92.7	-	-	-
eSDC+MSSDALF+DVR [57]	41.3	63.5	72.7	83.1	48.3	74.9	87.3	94.4	-	-	-
TDL [21]	56.3	87.6	95.6	98.3	56.7	80.0	87.6	93.3	-	-	-
AvgTAPR [24]	55.0	87.5	93.8	97.2	68.6	94.6	97.4	98.9	-	-	-
LOMO+KISSME+SRID [74]	65.5	85.4	91.3	95.7	83.0	95.3	97.5	99.3	-	-	-
LOMO+LMKDCCA [31]	73.3	90.5	94.7	98.1	86.4	97.5	99.6	**100**	-	-	-
CNN+RNN [22]	58.0	84.0	91.0	96.0	70.0	90.0	95.0	97.0	56.0	69.0	-
CNN+XQDA [19]	54.1	80.7	88.0	95.4	77.2	93.1	95.7	99.1	65.3	82.0	47.6
RFANet+RSVM [59]	49.3	76.8	85.3	90.0	58.2	85.8	93.4	97.9	-	-	-
DeepRCN+KISSME [76]	46.1	76.8	89.7	95.6	69.0	88.4	93.2	96.4	-	-	-
AMOC+ EpicFlow [30]	68.7	94.3	98.3	99.3	83.7	98.3	99.4	**100**	68.3	81.4	52.9
BRNN [77]	55.3	85.0	91.7	95.1	72.8	92.0	95.1	97.6	-	-	-
TRL [32]	57.7	81.7	-	94.1	87.8	97.4	-	99.3	80.5	91.8	69.1
TAM+SRM [26]	55.2	86.5	-	97.0	79.4	94.4	-	99.3	70.6	90.0	50.7
QAN [35]	68.0	86.6	95.4	97.4	90.3	98.2	99.3	**100**	73.7	84.9	51.7
ASTPN [36]	62.0	86.0	94.0	98.0	77.0	95.0	99.0	99.0	44.0	70.0	-
DSAN [29]	61.2	80.7	90.3	97.3	74.8	92.6	97.7	98.6	69.7	83.4	-
DRSTA [27]	80.2	-	-	-	93.2	-	-	-	82.3	-	65.9
CSSA+CASE [28]	85.4	96.7	98.8	99.5	93.0	99.3	**100**	**100**	**86.3**	94.7	76.1
RQEN [78]	76.1	92.9	97.5	99.3	92.4	98.8	99.6	**100**	73.7	84.9	51.7
STAL [33]	82.8	95.3	97.7	98.8	92.7	98.8	99.5	**100**	82.2	92.8	73.5
SDM [64]	60.2	84.7	91.7	95.2	85.2	97.1	98.9	99.6	71.2	85.7	-
SPL [65]	70.5	91.4	96.8	99.1	85.3	97.2	99.4	99.7	74.8	86.7	-
A3D (proposed)	**87.9**	**98.6**	**99.7**	**99.8**	**95.1**	**99.5**	**100**	**100**	**86.3**	**95.5**	**80.4**

The STFV3D [18] method first attempts to treat the pedestrian video as a 3D bin. Specifically, STFV3D segments the 3D bin into multiple spatial-temporal body-action sub-bins and uses fisher vector to aggregate them. However, STFV3D considers all sub-bins equally, ignoring that person matching may be misled by some “bad” sub-bins due to occlusions or clutter background. Compared with STFV3D, our proposed method further explores the salient parts in the 3D video bin with our A3D model. Compared with other methods also exploring spatial-temporal attentions, such as TAM+SRM [26], ASTPN [36], DSAN [29], and DRSTA [27], our A3D outperforms by a large margin, due to the consideration about constraint and interaction between temporal and spatial attention. CSSA+CASE [28] considers person ReID as a binary classification problem and devotes to distinguish whether given two video clips are the same identity, not preserving the rank relationship among multiple samples. This verification loss leads to a higher Rank-1 accuracy but relatively low mAP performance. Different from CSSA+CASE [28], we focus on the rank relationship in our objective function, and consequently, we obtain the same Rank-1 accuracy and improve mAP by 3.7% on MARS dataset, compared with CSSA+CASE [28]. Our A3D method also outperforms with a large margin over SDM [64] and SPL [65], which also introduce the RL to select key frames for video-based person ReID, since we additionally consider the spatial attention-aware learning in our A3D framework.

C. Ablation Study

To investigate the contribution of individual components in the A3D attention framework, we conducted two ablation evaluations on the iLIDS-VID dataset. First, we compared the complete A3D model with other incomplete settings, including (1) removing attention model; (2) replacing with a random attention agent; (3) removing the recurrent model; (4) removing the RL objective; and (5) remove the optical-flow stream. Second, we compared the proposed method with other attention methods with the same base-model and hyper-parameters, including (1) the QAN [35] method with our representation model and (2) the separate attention method which learns spatial attention on each frame first, and then learns the temporal aggregation attention using QAN.

Table VII summarizes the performance of the different variants of the proposed method. It is easy to draw the following conclusions from the comparison. (1) By comparing the baseline model without attention and our A3D model, we conclude that our A3D model can learn the salient bins of pedestrian video and improve the representation learning by the attention model. (2) The contrast between the proposed A3D agent and random attention agent shows that random attention agent without training predicts incorrect attention bins which mislead the identification. It indicates that the proposed A3D reinforcement learning framework learns a sensible policy for attention agent with the reward and hybrid loss. (3) As shown as ”w/o RNN” in Table VII, the performance drops without the RNN module, which demonstrates that the RNN is effective to obtain multiple attention clues. The motivation of the recurrent model is replacing the single attention bin with multiple ones for more robust representation learning. In the experiments, the RNN-based multi-head attention outperforms the single one with 1.1%/0.7% Rank-1/mAP improvement. It is not small since this improvement is achieved on a very strong baseline. (4) As shown as ”w/o RL objective” in Ta-

TABLE VII
ABLATION STUDIES ON THE ILIDS-VID DATASET.

Datasets	**iLIDS-VID**		
Rank@R	**R=1**	**R=5**	**mAP**
proposed	**87.9**	**98.6**	**92.5**
w/o attention	85.9	97.7	91.3
random attention	80.5	96.8	87.6
w/o RNN	86.8	98.5	91.8
w/o RL objective	85.0	96.6	90.1
w/o optical-flow	84.8	96.8	88.7
QAN*	85.8	98.1	91.0
Separate attention*	86.3	98.3	91.5

TABLE VIII
PARAMETERS ANALYSIS OF THE NUMBER OF STEPS IN THE RNN ON THE ILIDS-VID DATASET.

Datasets	**iLIDS-VID**		
Rank@R	**R=1**	**R=5**	**mAP**
one step	86.8	98.5	91.8
two steps	87.4	98.3	92.2
four steps (full)	**87.9**	**98.6**	**92.5**

ble VII, the performance is decreasing when we reduce the RL objective, which demonstrates the effectiveness and necessity of the RL algorithm. (5) The improvement with a large margin due to the optical-flow stream shows that the developed two-stream representation network effectively captures the motion information in the video and helps for person verification. AMOC+ EpicFlow [30] and CSSA+CASE [28] also utilize the optical flow as extra input to boost performance. (6) We have implemented and compared with the classic attention method QAN [35]. As shown as "QAN*" in Table VII, the proposed method also outperforms QAN* in terms of the accuracy at Rank=1 and mAP. The $*$ indicates that these methods are implemented using our base model with the same hyper-parameters. (7) The main difference between "separate attention" and QAN is that "separate attention" method employs the extra spatial attention. "separate attention" performing better than QAN indicates that the spatial attention is conducive to video representation. (8) The joint A3D method also improves over "separate attention*" which demonstrates that the constraints about spatial and temporal attention promote the attention learning.

D. Parameter Analysis

In this subsection, we analyzed the influences and sensitivity of major parameters. We conducted the parameters analysis experiments on the iLIDS-VID dataset. Specifically, we mainly investigate the number of glimpses, the size of attention region, the model complexity and show the convergence process.

Glimpse number: We further investigate how the performance of the proposed method changes when the number of steps in the RNN is varied. Table VIII summarizes the performance with different numbers of steps in the RNN model. The "step" denotes the number of recurrent cells in our RNN model, and indicates the number of attention bins. We observe that the performance is boosted correspondingly with the increasing of step numbers, since multiple attention maps can capture more discriminative information for person ReID. However, the improvement diminishes when the number of attention bins is enough. For example, the gap between the performance of two-step and four-step attention is faint.

TABLE IX
ANALYSIS OF TRAINING AND INFERENCE COST ON THE MARS DATASET.

Method	Training (h)	Inference (video/s)	Size (Mb)
w/o Attention	5.8	12.7	99.4
Our method	6.0	11.6	99.5

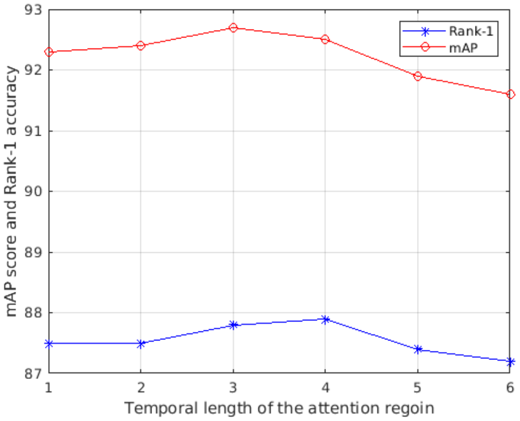


Fig. 5. Rank 1 accuracy and mAP scores on the iLIDS-VID dataset with different temporal lengths of attention region.

Attention region size: The size of the 3D attention region is also an important parameter of our method. We analysis the attention size in the both temporal and spatial domains. Fig. 5 shows the performance with different temporal lengths and fixed spatial size 6×3, which demonstrates that our A3D model is generally robust for the temporal lengths. However, too large temporal lengths reduce the performance, since the effect of temporal attention is weakened. As shown in Fig. 6, we also show the variances when the spatial size of attention is changed. The attention model with almost parameter settings achieve competitive results, which are better than 87%, except for too small attention regions. This may be because the small attention region is hard to locate the whole salient part.

Complexity analysis: As shown in Table IX, we show the computing costs of our method and the baseline (without attention) on the MARS dataset including training time, inference efficiency and model size. We can observe that our method requires negligible training and inference costs comparing to the baseline.

Convergence process: As shown in Fig. 7, we plot the training curves on different datasets where the abscissa denotes epoch number and ordinate denotes the testing performance. The blue and red curves respectively denote the Rank-1 and mAP metrics. All curves show that our method can achieve general convergence at 100 epochs. For the rest epochs, our method can gradually obtain better performance with some fluctuations.

E. Cross-Dataset Evaluation

In real surveillance systems, time and monetary cost are prohibitive to label overwhelming amount of data. Therefore, we usually apply our ReID system for unseen persons and scenes. In the conventional benchmarks, the pedestrian videos

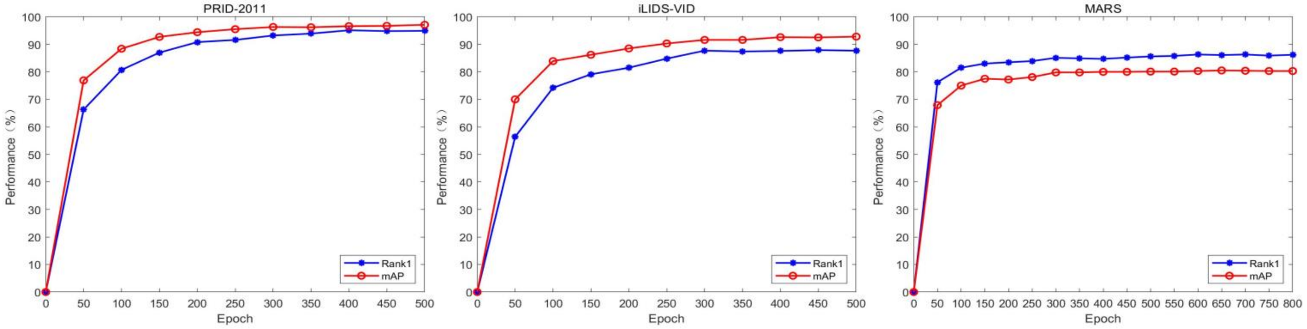


Fig. 7. The training processes on different datasets, which describe the development of performance with the increasing epochs. From left to right, we respectively display the curves of PRID-2011, iLIDS-VID, and MARS.

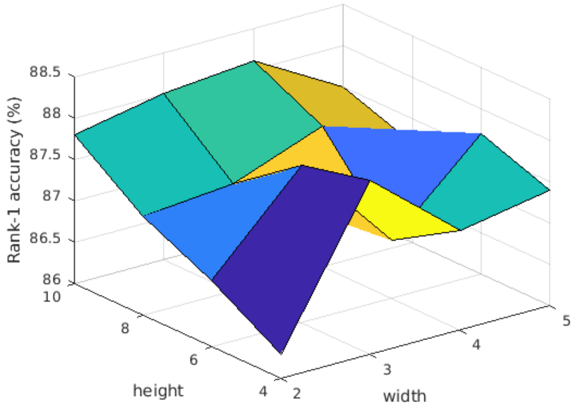


Fig. 6. Rank 1 accuracy on the iLIDS-VID dataset with different widths and heights.

TABLE X
RANK CMC ACCURACY OF CROSS-DATASET TESTING

Datasets	iLIDS-VID/PRID-2011			
Rank@R	**R=1**	**R=5**	**R=10**	**R=20**
CNN+RNN [22]	28.0	57.0	69.0	81.0
QAN [35]	34.0	61.3	74.0	83.1
ASTPN [36]	30.0	58.0	71.0	85.0
RQEN [78]	61.8	82.6	90.4	96.1
STAL [33]	63.7	84.0	**92.8**	98.1
A3D (proposed)	**64.2**	**84.7**	**92.8**	**98.5**

in both training and testing processes are captured from the same surveillance scenarios, where illumination conditions and backgrounds are relativity consistent. Even though some transfer learning based methods have been prospoed to transfer the model from source domain to the target domain, they always need the information (e.g. images or attributes) of target domain and train the target-specific model with these information, which are not immediately available. To evaluate the effects of the proposed method applied to a real-world surveillance system, we conducted the *cross-dataset testing* [22], where the model is trained by a randomly split training set of the iLIDS-VID dataset and tested on the testing set of the PRID-2011 dataset. Besides, we also repeated the evaluations for 10 splits and calculated the average accuracy.

Table X summarizes the performance of the proposed method and the state-of-the-art methods in the cross-dataset testing, including CNN+RNN [22], QAN [35], ASTPN [36], RQEN [78], and STAL [33]. In the cross-dataset setting, we also obtain Rank 1 accuracy improvement compared with other methods. Compared with CNN+RNN [22], ASTPN [36] and QAN [35], we achieve a significant improvement with spatial-temporal attention model. While compared with RQEN [78] and STAL [33], we still slightly outperform them by the 3D consistent.

F. Visualization of Attention Model

In Figure 8, we depict the 3D attention regions generated by the proposed A3D method of 3 example persons in the iLIDS-VID dataset. In our A3D model, the agent iteratively predicts 4 attention bins for given video clips. We select the last three attention to show in Figure 8 as purple, blue and orange cuboids respectively. Figure 8 (a) shows the complete recurrent 3D attention bins in the original prediction video clips. While Figure 8 (b) displays the specific spatial attention region location for the third frame of each video. With the visualization of attention bins on pedestrian video examples, we demonstrate the proposed recurrent A3D attention model precisely captures the salient spatial regions and key temporal frames. Specifically, the top example shows that the A3D attention selects the frames with less occlusions, and the bottom example indicates the proposed agent tends to select the spatial regions without occlusions. The middle example shows that the A3D agent locates the more salient part for the video without occlusions. Besides, for the salient parts of the video clips, our A3D agent even favors more glimpses on the same location.

V. CONCLUSIONS

In this work, we have proposed a recurrent 3D attention reinforcement learning framework to consider the spatial-temporal consistency in the attention detector. In our framework, we first extract both appearance and motion information with a two-stream network and then apply an RNN model to iteratively select multiple salient 3D bins in the video. To train the non-differentiable attention selection, we formulate it as MDP and optimize it with REINFORCE. We evaluated our method on three video person re-identification datasets and demonstrated consistent improvement of our proposed approach over state-of-the-art methods. The attention bins in our approach are rigid, which are restricted by the decision

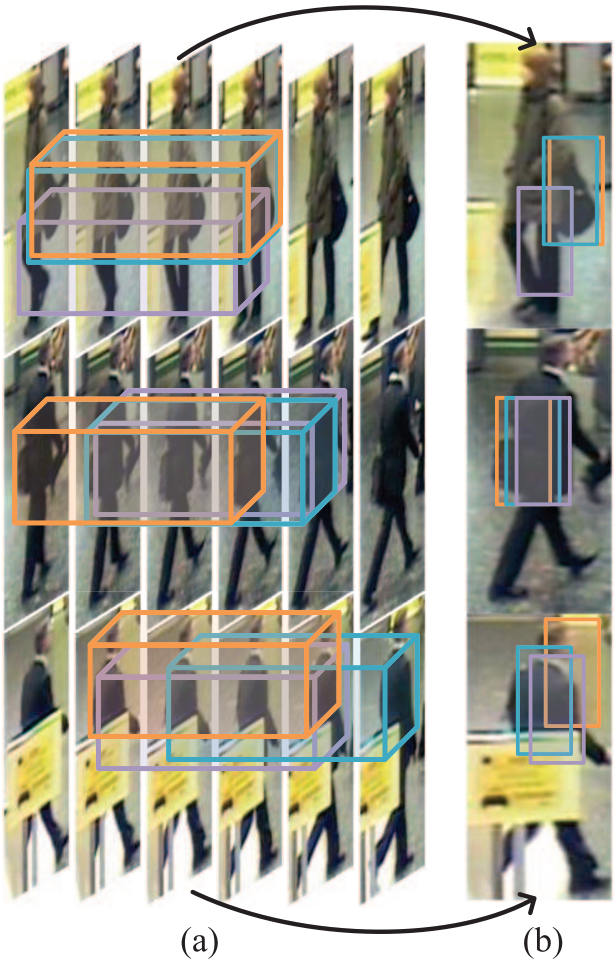


Fig. 8. 3D attentions of person samples in the iLIDS-VID dataset. (a) shows the recurrent 3D attention bins predicted by the proposed method in the original prediction video clips. To precisely visualize the locations of spatial attention regions, we draw the third frame of each video clip as examples and display the spatial regions in these frames in the part (b). Taking the first clip as the example, the part (a) of this figure shows that the learned attention bins focus on the frames with less occlusions from signboard, while the part (b) demonstrates that these attention bins focus on the salient spatial region (the bag). Best viewed in color.

complexity of an attention agent. In the future, we will try to learn a deformable 3D attention bin, which is more flexible and 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] G. Chen, T. Zhang, J. Lu, and J. Zhou, “Deep meta metric learning,” in *ICCV*, pp. 9547–9556, 2019.

[3] G. Chen, C. Lin, L. Ren, J. Lu, and J. Zhou, “Self-critical attention learning for person re-identification,” in *ICCV*, pp. 9637–9646, 2019.

[4] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, Q. Tian, *et al.*, “Person re-identification in the wild.,” in *CVPR*, vol. 1, p. 2, 2017.

[5] 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.

[6] F. Xiong, M. Gou, O. Camps, and M. Sznaier, “Person re-identification using kernel-based metric learning methods,” in *ECCV*, pp. 1–16, 2014.

[7] 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.

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

[9] 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.

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

[11] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, “Beyond part models: Person retrieval with refined part pooling,” in *ECCV*, 2018.

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

[13] Z. Zhang, Y. Xie, W. Zhang, Y. Tang, and Q. Tian, “Tensor multi-task learning for person re-identification,” *IEEE Transactions on Image Processing*, 2019.

[14] Z. Zhang, Y. Xie, W. Zhang, and Q. Tian, “Effective image retrieval via multilinear multi-index fusion,” *IEEE Transactions on Multimedia*, vol. 21, no. 11, pp. 2878–2890, 2019.

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

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

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

[18] 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.

[19] 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.

[20] 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.

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

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] S. Li, S. Bak, P. Carr, and X. Wang, “Diversity regularized spatiotemporal attention for video-based person re-identification,” in *CVPR*, pp. 369–378, 2018.

[28] D. Chen, H. Li, T. Xiao, S. Yi, and X. Wang, “Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding,” in *CVPR*, pp. 1169–1178, 2018.

[29] L. Wu, Y. Wang, J. Gao, and X. Li, “Where-and-when to look: Deep siamese attention networks for video-based person re-identification,” *TMM*, 2018.

[30] 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.

[31] 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.

[32] J. Dai, P. Zhang, D. Wang, H. Lu, and H. Wang, “Video person re-identification by temporal residual learning,” *TIP*, 2018.

[33] G. Chen, J. Lu, M. Yang, and J. Zhou, “Spatial-temporal attention-aware learning for video-based person re-identification,” *IEEE Transactions on Image Processing*, 2019.

[34] X. Lan, H. Wang, S. Gong, and X. Zhu, “Deep reinforcement learning attention selection for person re-identification,” *arXiv preprint arXiv:1707.02785*, 2017.

[35] Y. Liu, J. Yan, and W. Ouyang, “Quality aware network for set to set recognition,” in *CVPR*, 2017.

[36] 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.

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

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

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

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

[41] 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.

[42] 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*, 2017.

[43] 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.

[44] 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.

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

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

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

[48] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *JMLR*, vol. 10, no. Feb, pp. 207–244, 2009.

[49] 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.

[50] A. Mignon and F. Jurie, "Pcca: A new approach for distance learning from sparse pairwise constraints," in *CVPR*, pp. 2666–2672, IEEE, 2012.

[51] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith, "Learning locally-adaptive decision functions for person verification," in *CVPR*, pp. 3610–3617, 2013.

[52] S. Bai, X. Bai, and Q. Tian, "Scalable person re-identification on supervised smoothed manifold," in *CVPR*, 2017.

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

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

[55] S. Liao and S. Z. Li, "Efficient psd constrained asymmetric metric learning for person re-identification," in *ICCV*, pp. 3685–3693, 2015.

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

[57] 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.

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

[59] 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.

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

[61] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *CVPR*, p. 2, 2018.

[62] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Mask-guided contrastive attention model for person re-identification," in *CVPR*, pp. 1179–1188, 2018.

[63] C. Lin, J. Lu, G. Wang, and J. Zhou, "Graininess-aware deep feature learning for pedestrian detection," in *ECCV*, pp. 732–747, 2018.

[64] J. Zhang, N. Wang, and L. Zhang, "Multi-shot pedestrian re-identification via sequential decision making," in *CVPR*, 2018.

[65] D. Ouyang, J. Shao, Y. Zhang, Y. Yang, and H. T. Shen, "Video-based person re-identification via self-paced learning and deep reinforcement learning framework," in *ACM MM*, pp. 1562–1570, 2018.

[66] Q. Cao, L. Lin, Y. Shi, X. Liang, and G. Li, "Attention-aware face hallucination via deep reinforcement learning," in *CVPR*, pp. 690–698, 2017.

[67] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *NIPS*, pp. 568–576, 2014.

[68] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in *ICCV*, pp. 2758–2766, 2015.

[69] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine learning*, vol. 8, no. 3-4, pp. 229–256, 1992.

[70] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, pp. 770–778, 2016.

[71] 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.

[72] 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.

[73] Y. Li, Z. Wu, S. Karanam, and R. J. Radke, "Multi-shot human re-identification using adaptive fisher discriminant analysis," in *BMVC*, pp. 1–12, 2015.

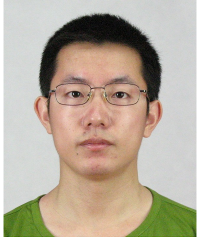
[74] S. Karanam, M. Gou, Z. Wu, A. Rates-Borras, O. Camps, and R. J. Radke, "A systematic evaluation and benchmark for person re-identification: features, metrics, and datasets," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 3, pp. 523–536, 2018.

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

[76] 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.

[77] W. Zhang, X. Yu, and X. He, "Learning bidirectional temporal cues for video-based person re-identification," *TCSVT*, 2018.

[78] G. Song, B. Leng, Y. Liu, C. Hetang, and S. Cai, "Region-based quality estimation network for large-scale person re-identification," in *AAAI*, 2018.



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 (Senior Member, IEEE) received 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. His current research interests include computer vision, pattern recognition, and intelligent robotics, where he has authored/co-authored over 250 scientific papers in these areas which have been cited over 10,000 times. He was/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 serves as the Co-Editor-of-Chief for Pattern Recognition Letters, an Associate Editor for 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 also serves as the Program Co-Chair of IEEE FG'2023, IEEE VCIP'2022, IEEE AVSS'2021 and IEEE ICME'2020.

Ming Yang (Member, IEEE) 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 80 peer reviewed publications in prestigious international journals and conferences, which have been cited over 13,000 times. He is a member of the IEEE.

Jie Zhou (Senior Member, IEEE) 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.