IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 7, 2018, accepted December 25, 2018, date of publication January 1, 2019, date of current version January 23, 2019.

Digital Object Identifier 10.1109/ACCESS.2018.2890394

Recurrent Models of Visual Co-Attention for Person Re-Identification

LAN LIN[1], HUAN LUO[1], RENJIE HUANG[2], AND MAO YE[1]

[1]Center for Robotics, Key Laboratory for NeuroInformation of Ministry of Education, School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China
[2]School of Computer Science and Information Science, Southwest University, Chongqing 400715, China

Corresponding author: Mao Ye (cvlab.uestc@gmail.com)

This work was supported in part by the National Natural Science Foundation of China (61773093), Key R&D Program (Intelligent Processing Technology of Multi-source Litigation Letters and Visits National 2018YFC0831800), Important Science and Technology Innovation Projects in Chengdu (2018-YF08-00039-GX) and Research Programs of Sichuan Science and Technology Department (2016JY0088, 17ZDYF3184).

ABSTRACT Person re-identification (re-id) refers to matching people across disjoint camera views. Most of person re-id methods extract discriminative features from the whole images or fixed regions and develop their metrics. However, these methods ignore that the attention regions with temporal cues in the pedestrian image pair hold discriminative information. In this paper, we propose the recurrent models of visual co-attention that aim to simulate human eye movement, focusing on the sequential concurrent attention (co-attention) regions of the same locations when comparing image pairs. Since reinforcement learning provides a flexible learning strategy for sequential decision-making, it is naturally applied to perform the temporal re-id co-attention learning task. The reward functions are designed to recursively optimize the prediction by rewarding or punishing the learning process. The recurrent models are used to extract information from a sequence of attention regions. Finally, person re-id is performed based on the whole image feature and the features from the recurrent models. Our contributions are: 1) the visual mechanism, which can dynamically locate the optimal co-attention regions to simulate the human re-id process; 2) the design of reward functions in reinforcement learning, which aims to recursively optimize the prediction process; and 3) experimental results, which demonstrate the advantages of our method compared with the state-of-the-art methods.

INDEX TERMS Person re-identification, eye movement, reinforcement learning, recurrent neural network, co-attention mechanism.

I. INTRODUCTION

Person re-id aims to recognize all persons from a gallery who have the same identity to the probe. The gallery persons and the probe persons are captured from different non-overlapping camera views across temporal periods. Since person re-id is widely applied in intelligent video surveillance, criminal investigation, and long-term pedestrian tracking, it becomes increasingly important. However, person re-id remains a challenging problem due to significant intra-class variations of illumination, view angle, pedestrian pose, and occlusion. To address these challenges, many methods have been proposed which can be roughly classified into two categorizations: feature-based methods and metric-based methods.

For feature-based methods, significant efforts are devoted to extract robust handcrafted features [1]–[4] or learn discriminative deep features [5]–[11]. Although these feature-based methods have achieved encouraging performances, they are extracted from the whole images or fixed regions. Thus the features would be misaligned and not be robust when occlusions or various poses occur.

As for metric learning-based methods, they are developed to maximize the inter-class feature distance and minimize the intra-class distance by projecting the raw data to the learned metric space [1], [12]–[16]. However, since the pedestrian appearances undergo large variation across multi-camera views, there would be still some confusing negative gallery pedestrians who tend to be more similar to a probe than positive ones even after metric space learning.

Both feature-based methods and metric learning-based methods pay more attention to the whole images and fixed regions, which fail to match well in the case of large appearance variation and occlusion. In fact, some researchers have realized that attention regions with temporal cues

FIGURE 1. The illustration shows the motivation of our method. Different pedestrian image pairs should hold various temporal co-attention processes due to dynamic pairwise image environment. The top left orange box indicates positive sample, and the other three indicate negative samples.

in pedestrian images hold discriminative information in re-id [17]–[21]. These methods are based on recurrent neural network (RNN). RNN is a type of deep neural network that has recurrent connections, which enables the network to extract information in the sequence and memorize the internal states [22], [23]. However, these methods do not consider exploiting reinforcement learning (RL) based on RNN to choose attention regions in person re-id, while RL is a problem dealt by an agent that learns its optimal behavior by trial-and-error interactions with a dynamic environment [24], [25], providing a flexible learning strategy for sequential decision-making. In the process of comparing two images, human focus on a series of identical attention regions through repeatedly looking back and forth (i.e., co-attention by eye movements), which is actually a sequential decision-making process interacting with a dynamic pairwise image environment. Therefore it is natural to use RL to imitate the scan order of our eyes, carrying out the temporal re-id co-attention learning process. This process is in accordance with the process of human vision perception [26]–[28] and is subtly robust to occlusion and large pose variation. Moreover, the co-attention regions and the whole image pair hold distinct intrinsic information of multiple scales. That is to say, for a given pair, human not only see the details of local regions, but also get an impression on the whole images. The examples of pedestrian image pairs with various temporal co-attention processes are shown in Fig. 1. We can see that, even for image pairs composed of the same probe and different galleries, the locations and scales on which humans fixate should be strongly task-specific. It is meaningful to learn to adaptively co-concentrate on the appropriate local regions and visual scales over time for different image pairs, and use the past information to guide what and where to look at next.

In this paper, we propose Recurrent Models of visual Co-Attention (RMCA) for person re-id motivated to simulate the co-attention process of eye movements. The proposed models consider the co-attention mechanism as the sequential decision process of a task-driven agent interacting with a pairwise image environment. The architecture of RMCA is shown in Fig. 2. First, given an input image pair and the co-attention coordinates l^{t-1} of the glimpse, the local features centered at the coordinates are extracted. Next, local features are combined with the internal representations memorized by the hidden layer of RNN, and new internal representations are generated. Then, the generated internal representations are used to learn a location policy on the basis of re-id pairwise label constraints, which decides where and what to co-attend to in next glimpse. This local co-attention matching based on RNN iteration is repeated for several time steps. Finally, the deep global features are extracted at the end of the time sequence, and they are concatenated with the internal representations. The joint features are used to learn the identification action and triplet ranking action by the designed reward functions of RL. In the architecture, the sequence of co-attention location generation simulates the scan order of our eyes. Recurrent feature extraction and identification/triplet ranking action in RL imitate the function of our visual system and our brain respectively. As far as we know, the proposed RMCA is the first attempt to exploit reinforcement learning to choose an attention region sequence in person re-id.

The main contributions of our work are as follows:

(1) We propose recurrent models of visual co-attention for person re-id called RMCA. Our model is able to dynamically locate the optimal co-attention regions, and combine them with the global matching into an integrated framework for simulating the human re-id process.

(2) We formulate the sequential co-attention matching in RMCA as a reinforcement learning based on

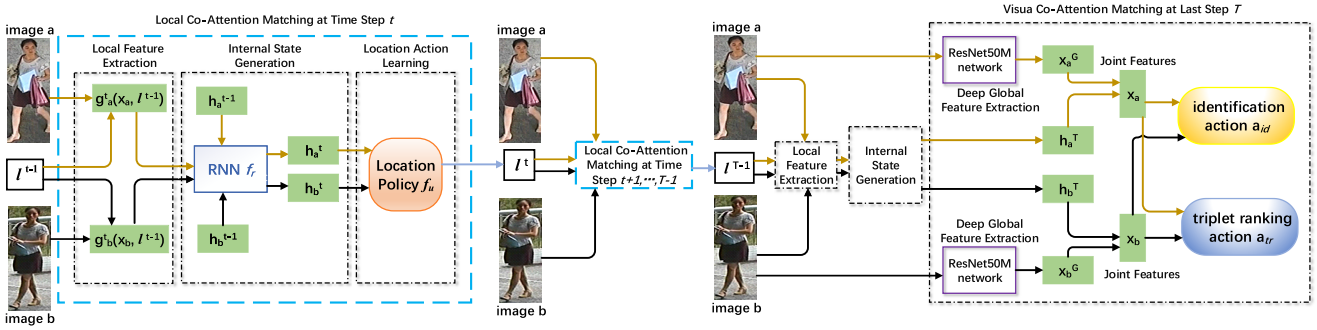


FIGURE 2. The architecture of the proposed RMCA. RNN is used to extract features from a sequence of local attention regions. These representations are used to learn a location policy on the basis of re-id pairwise label constraints, which decides the next co-attention locations. Finally, the global features are combined with the internal representations of recurrent models. Person re-id is performed based on these joint features.

RNN architecture. And the reward functions of RL are designed to recursively optimize the co-attention locations and the interactive sequence information for prediction by rewarding or punishing the learning process.

(3) Extensive comparisons are conducted between RMCA and the state-of-the-art models on Market-1501 [29], CUHK03 [5], and CUHK01 [30] person re-id benchmarks, demonstrating the superior performance of our method.

The remainder of this paper is organized as follows. Section II gives a brief introduction of the existing methods. Section III describes the details of our recurrent models of visual co-attention based on reinforcement learning. Section IV reports the experimental results, comparing our method with the state-of-the-art methods. Section V makes some conclusions.

II. RELATED WORKS

Many existing person re-id methods try to extract handcrafted features or learn deep features across cameras. The handcrafted features are technically designed to be robust to the person appearances across cameras by alleviating the effect of various pose, viewpoint, and illumination [1]–[3]. Liao *et al.* [1] analyze the horizontal stripes and utilize the local maximal occurrence (LOMO) feature against viewpoint changes. Matsukawa *et al.* [2] propose a region descriptor based on hierarchical Gaussian distribution of pixel features, which includes the mean and covariance formation. In [3], a multiple hypergraph fusion (multi-HG) method is employed to extract complementary information from different feature descriptors. And with the great success of deep learning in many computer vision tasks, the deep learning based features are widely adopted in person re-id recently [5], [31], [32]. Li *et al.* [5] propose a filter pairing neural network (FPNN) that introduces a patch-matching layer in the convolutional neural network(CNN), which can handle the part displacement in each horizontal stripe. Chen *et al.* [31] formulate a deep pyramid feature learning (DPFL) CNN architecture to handle multi-scale feature fusion. Jiao *et al.* [32] propose a deep model for joint learning of image super-resolution and person identity classification.

Besides robust features, metric learning has been widely applied to dealing with the complex matching problem for person re-id. Cross view quadratic discriminant analysis (XQDA) [1] automatically selects the optimal dimensionality as a discriminative subspace, and learns its distance metric simultaneously. Liao and Li [33] propose to weight the positive and negative samples differently with a positive semidefinite constraint. In recent re-id methods based on CNN structure, the identification loss and the triplet loss are widely adopted. The identification loss [34] is emerging with the advent of large datasets, which does not require random sampling. Hermans *et al.* [35] propose an improved triplet loss, which can do online sample selection within each batch.

Recently, feature sequence based methods such as recurrent feature aggregation and attention mechanism have been adopted in person re-id [17], [18], [21], [23], [36]. In [23] and [36], local features of recurrent appearance data are extracted and aggregated using RNN, while they still suffer from occlusion and large pose variation problem. Xu *et al.* [21] present a joint spatial and temporal attention pooling network, which is just applied to video-based person re-id. Liu *et al.* [18] propose an attention model which generates and integrates a series of different local parts by masking CNN feature map, while the masked feature map is not quite consistent with human perception. Si *et al.* [17] learn intra-sequence feature refinement and inter-sequence feature-pair alignment via a dual attention mechanism, though there is no correlation across the learned sequences.

There has been a few previous work incorporating the idea of reinforcement learning in computer vision tasks. Some that share the same spirit as our work include image classification [27], object localization [37], [38] and active object recognition [39]. In person re-id, Lan *et al.* [40] formulate an identity attention as reinforcement learning model for cropping given auto-detected bounding boxes, which is more about post-detection than the sequence of attention regions.

III. RECURRENT MODELS OF VISUAL CO-ATTENTION

In this section, we first introduce problem formulation. Second, the proposed recurrent co-attention matching based on reinforcement learning is explained with more details. Third, model training is displayed to optimize the sequential actions. Finally, the test procedure is described.

A. PROBLEM FORMULATION

The motivation of RMCA is that, when comparing two images, human focus attention selectively on a series of local regions to acquire and combine information through eye movements. And the guidance of human eye movements is based on past pairwise information and the demands of pedestrian matching. To mimic the aforementioned human vision perception in person re-id, we formulate the local co-attention matching as a sequential decision process through reinforcement learning built on RNN architecture. RNN processes pairwise inputs sequentially and combines information over time to build up a dynamic internal representation of environment. The task-specific agent in RL interacts with such dynamic environment and thus decides the next co-attention regions with the purpose of maximizing the reward.

At each time step, via observing the environment, the agent extracts local features of the identical location from a pair of images. Let g_a^t and g_b^t denote the local features at t-th time step captured from two images x_a and x_b, respectively. In the RNN based model, g_a^t and g_b^t are combined with the internal representations at previous time step h_a^{t-1} and h_b^{t-1}, respectively, and the new internal states of the model, h_a^t and h_b^t, are generated. The location l^{t-1} and the matching action of g_a^t and g_b^t are determined by the past local observations h_a^{t-1} and h_b^{t-1}. Similarly, h_a^t and h_b^t would affect the next location l^t to attend to and match. The purpose of the agent is to design a task-specific reward function R and is capable of adaptively

selecting a sequence of discriminative regions by carrying out the action policy. Additionally, the agent needs to maximize the cumulative rewards of all time steps, and it may sacrifice immediate reward to gain more long-term reward, thus the agent can integrate information over time most effectively.

B. CO-ATTENTION MATCHING BASED ON REINFORCEMENT LEARNING

As shown in Fig. 2, the proposed model exploit reinforcement learning based on RNN architecture to simulate local co-attention matching, with the purpose of locating the optimal regions and improving re-id performance. To this end, the states, actions and rewards are defined as follows.

States: At each step t, given an image pair (x_a, x_b) and the location l^{t-1} of co-attention region, the agent extracts local features $g_a^t(x_a, l^{t-1})$ and $g_b^t(x_b, l^{t-1})$, respectively. The internal representations h_a^{t-1} and h_b^{t-1} are memorized by the hidden layer of RNN, which keeps the history of past observations. The states h_a^t and h_b^t are updated by mapping the local features and internal representations into hidden spaces:

$$h_a^t = f_r(w_a^h h_a^{t-1} + w_a^g g_a^t + b_a^h), \tag{1}$$
$$h_b^t = f_r(w_b^h h_b^{t-1} + w_b^g g_b^t + b_b^h), \tag{2}$$

where w_a^h and w_a^g are the forward weights of hidden layer and input layer for image x_a, respectively; w_b^h and w_b^g are the forward weights of hidden layer and input layer for image x_b, respectively; b_a^h and b_b^h are bias parameters. The activation function is $f_r(x) = \max(x, 0)$, which is the rectified linear unit (ReLU) function. The concatenation h_{conca}^t of h_a^t and h_b^t, i.e., $h_{conca}^t = [h_a^t; h_b^t]$, summarizes information extracted from past observations and is sufficient to determine where to co-attend to next. Therefore, h_a^t and h_b^t can also be referred to as Markov states.

Actions: The agent in this work executes three actions: the location action a_l, the identification action a_{id} and the triplet ranking action a_{tr}.

The location action a_l is defined to assist the agent to determine the next co-attention location. At each step t, the agent would generate a mean value μ^t based on the history of past observations h_{conca}^t:

$$\mu^t = f_\mu(w_l h_{conca}^t + b_l), \tag{3}$$

where w_l and b_l are the weights and bias of coordinate calculation, respectively; h_{conca}^t is the concatenation of internal states h_a^t and h_b^t. The activation function is $f_\mu(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$. Gaussian distribution is generated by the mean μ^t and a fixed standard variance σ, and its probability density function (PDF) is shown as follows:

$$f(x|\mu^t, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu^t)^2}{2\sigma^2}\right). \tag{4}$$

A noise value x is stochastically choosen from the above Gaussian distribution, and x is combined with the mean μ^t to determine the next co-attention location:

$$l^t = \mu^t + x. \tag{5}$$

It is worth noting that the co-attention location l is generated by a two-component Gaussian, and the two components correspond to the horizontal coordinate and vertical coordinate, respectively. Eq.(4) defines the PDF of Gaussian distribution for each component. In order to describe the inference procedure conveniently, we only describe the generation of one coordinate and use l to represent the location. A small variance value is chosen as $\sigma = 0.17$, since the range of eye movement at each step would not likely be drastic.

The agent applies the identification action a_{id} to learn from the image-identity correspondence relations and use the learned knowledge for matching other unseen pedestrians [31]. The local attention features $X = \{x_i\}_{i=1}^n$ are extracted from a mini-batch of n images, and the identity class labels of these images are denoted as $Y = \{y_i\}_{i=1}^n$. These training images capture m_{id} different persons (i.e. $y_i \in [1, \cdots, m_{id}]$). Consequently, the identification action a_{id} adopts a softmax output:

$$p_i = p(y = y_i|x_i) = \frac{\exp(w_i x_i + b_i)}{\sum_k \exp(w_k x_k + b_k)}, \tag{6}$$

where w_k and b_k refer to the identification weight and bias of the training identity class k.

The triplet ranking action a_{tr} is utilized to mine hard samples for learning critical information. An anchor sample x_{a_j} is randomly selected from camera A in the mini-batch of m_{tr} anchor samples (i.e. $j \in [1, \cdots, m_{tr}]$), and the triplet ranking action a_{tr} of x_{a_j} is performed as follows:

$$z_{a_j} = m + \max d(x_{a_j}, x_b^+) - \min d(x_{a_j}, x_b^-), \tag{7}$$

where x_b^+ and x_b^- respectively represent all positive and negative samples of x_{a_j} in this mini-batch from camera B; $d(x_{a_j}, x_b^+)$ and $d(x_{a_j}, x_b^-)$ indicate the Euclidean distance of positive image pairs and negative image pairs, respectively. The agent carries out the action a_{tr} to mine the hardest positive sample and the hardest negative sample through ranking $\max d(x_{a_j}, x_b^+)$ (i.e., maximum distance) and $\min d(x_{a_j}, x_b^-)$ (i.e., minimum distance) with a margin value $m = 0.3$. This is in the spirit of online triplet selection [18], [35]. The image pairs selected within a mini-batch perform content-aware actions since their concatenation h_{conca}^t is dependent on pairwise environment. Specifically, in the sequential decision process with T time steps, the location action a_l is executed in the first $T-1$ steps, while at the last step T, CNN is utilized to learn global discriminative features concatenated with internal representations later. The joint feature vectors are exploited to perform the identification action a_{id} and the triplet ranking action a_{tr} at step T.

Rewards: The reward function is defined as $R = \sum_{t=1}^T r^t$. At each step t, based on the history of past interactions with the environment, h_a^{t-1} and h_b^{t-1}, the agent needs to learn co-attending to the appropriate local regions (g_a^t, g_b^t) of image pair (x_a, x_b), which is subject to re-id matching criterion. And then the agent takes the above actions a_l, a_{id} and a_{tr}. If the actions serve long term interests such as selecting the discriminative local regions or keeping away from occlusions,

they should be encouraged, otherwise they should be punished. Thus, r_t is defined to get a scalar feedback signal to fulfill the encouragement or punishment in each step t as follows:

$$r^t = \sum_{i=1}^{m_{id}} I(y = y_i) + \sum_{j=1}^{m_{tr}} I(z_{a_j} \leqslant 0), \tag{8}$$

where y and y_i are respectively the prediction value and identity label in Eq.(6), z_{a_j} is the result of Eq.(7), and $I(x)$ is the indicator function, i.e., $I(x)$ equals to 1 when x is true and 0 otherwise. Intuitively, the term $I(y = y_i)$ commits a positive reward if the correspondence relation between attention region and identity is proper, or no reward otherwise. The other term $I(z_{a_j} \leqslant 0)$ encourages the true rank of matching in the selected triplets from one batch. Finally, the rewards of all m_{id} identities and m_{tr} triplets within a mini-batch are accumulated to teach the agent to focus its attention on the discriminative regions.

C. HYBRID OBJECTIVE FUNCTION

The parameters in the location training are learned so as to maximize the total reward. Most parameters just involve the gradients of the CNN (feature extraction) and RNN (observation memory), which can be computed by standard backpropagation [41]–[43]. However, our co-attention location generation adopts a non-differentiable stochastic unit with the condition input, while reinforcement algorithm is powerful to optimize the stochastic unit by a sample approximation to the gradient [27], [44]. It is formulated as:

$$\frac{\partial r^t}{\partial \mu^t} = (r^t - b^t)\frac{x - \mu^t}{\sigma^2}, \tag{9}$$

where r^t is calculated from Eq.(8); μ^t is the mean of the co-attention locations in Eq.(3) and is also the input of reinforcement learning; $f(x|\mu^t, \sigma^2)$ is shown in Eq.(4), which is the PDF of a two-component Gaussian distribution; x is the random sampled value; σ is set to a fixed standard deviation; b^t is a baseline value to reduce the variance of the gradient of r^t w.r.t. μ^t [27], [45], and in this work b^t is the linear transformation of h^t_{conca}. b^t is learned via $l_b = (b^t - r^t)^2$. By making b^t a moving average of r^t, the baseline b^t learns in effect at the same rate as the rest of the model.

Furthermore, the correct label associated with a training sample is known at the end of an observation sequence. Then in a mini-batch the cross entropy loss l_{id} [31], [46] is utilized to train the identification action a_{id}, and the triplet loss l_{tr} [35] is applied to train the triplet ranking action a_{tr}. l_{id} and l_{tr} are shown as follows:

$$l_{id} = -\sum_{i=1}^{m_{id}} (y_i \ln(p_i) + (1 - y_i)\ln(1 - p_i)), \tag{10}$$

$$l_{tr} = \sum_{i=1}^{m_{tr}} [m + \max d(x_{a,i}, x^+_{b,i}) - \min d(x_{a,i}, x^-_{b,i})]_+. \tag{11}$$

where $[\cdot]_+$ truncates the involved variable at zero. Finally, a hybrid objective function is constructed to optimize actions a_l, a_{id} and a_{tr} as follows:

$$l = \lambda_1 l_{id} + \lambda_2 l_{tr} + \lambda_3 l_b - \lambda_4 R, \tag{12}$$

where $\lambda_1 = 1$, $\lambda_2 = 1$, $\lambda_3 = 0.5$ and $\lambda_4 = 1$ are hyperparameters. The hybrid loss is optimized to train the recurrent models and backpropagate the gradients through the models. It is worth nothing that the gradients of location training are always learned by the sample approximation techniques from reinforcement algorithm [27], [44].

D. TEST PHASE

In the test phase, the learned co-attention network RMCA is applied to all test pedestrian image pairs. The test process mainly includes three steps. First, given any test pedestrian image pair and the co-attention coordinates learned by the location policy, the sequential co-attention regions are generated. Then, the learned recurrent models are used to extract features from the sequential co-attention regions and produce internal representations over time. At last, a joint feature vector is obtained by concatenating internal representation and the deep global feature vector extracted at time step T. The joint feature vector is used to compute Euclidean distance between such image pair.

IV. EXPERIMENTS

A. IMPLEMENTATION DETAILS

The proposed RMCA are implemented in the PyTorch framework. The main configuration of our computer is 2 × NVIDIA GeForce GTX 1080 Ti GPUs. Before training the RMCA model, we pre-train the base network ResNet50M [42] with the standard softmax classification on the ImageNet dataset [47] for model initialization. After pre-training, we remove the last two layers and output 3072-dimension feature vector of the pre-trained ResNet50M network to initialize the global CNN part of our model introduced in Section III-B. The dimension of the RNN hidden layer is set to 1024. In all of experiments, pedestrian images are resized to 256 × 128 in pixel, while the size of local co-attention regions is set to 112 × 56. The local feature g^t is obtained by linear transformation, which is simple but effective. Certain complex CNN feature extraction may be valuable but inevitably involves in increasing computation and complexity. The time step T of RNN is set to 5. The model is trained by 240 epochs with stochastic gradient descent (SGD) algorithm. The learning rate is initialized as 0.003 and changed to 0.0003 in the last 120 epochs.

In addition, to overcome the imbalance issue of positive pairs and negative pairs in the training set of triplet ranking action, we augment data by following the common techniques such as random translational transforms [5], random horizontal flips and random crops [48] on each pedestrian image. The triplet mini-batches are generated through the online triplet mining strategy [35], which can produce more triplets by randomly shuffling the dataset according to identity labels. Each mini-batch includes sixteen persons/identities, and each of them has four images.

TABLE 1. The experimental setting on three person re-id datasets. The person bounding box split on CUHK03 only shows the first random split. SQ: Single-Query; MQ: Multi-Query; SS: Single-Shot; MS: Multi-Shot.

Dataset	No. of identities	Identity Split		Person Bounding Box Split			Test Setting	No. of random splits
		Training	Test	Training	Gallery	Probe		
Market-1501	1,501	751	750	12,936	19,732	3,368	SQ, MQ	–
CUHK03	1,467	1,367	100	13,132	489	475	SS	20
CUHK01	971	485	486	970	972	972	MS	10

(a) (b) (c) (d)

FIGURE 3. Examples of pedestrian image pairs from (a) Market-1501 dataset, (b) CUHK03 dataset with manually cropped person images, (c) CUHK03 dataset with automatically detected bounding boxes, and (d) CUHK01 dataset.

B. COMPARISON WITH STATE-OF-THE-ART METHODS

1) EXPERIMENTS ON MARKET-1501

RL needs large data for model training. Market-1501 [29] and CHHK03 [5] is currently the largest and the second largest public available image-based re-id datasets. Thus, it is natural to choose these two datasets for our models. Market-1501 [29] is a realistic person re-id dataset captured from six camera views of different resolutions in front of a campus supermarket at Tsinghua University. It contains 32,668 bounding box images of 1,501 pedestrians. Fig. 3(a) shows several examples of Market-1501 bounding box images. The standard split setting [29] is adopted, i.e., 751 pedestrians are used for training and 750 for test. The training set includes 12,936 images. The test set contains 3,368 query images and 19,732 gallery images with junk and distractor ones. The training/test data splits and testing settings of each dataset is summarized in Table 1. The cumulative matching characteristic (CMC) curve is employed to measure the performance of the proposed method on Market-1501 dataset and the following two datasets. Besides, mean Average Precision (mAP) is also adopted in Market-1501 to compute the mean of average precision over all probes, since there is an average of 14.8 cross-camera ground truth matches for each query.

The proposed RMCA is compared with fourteen state-of-the-art methods under both the single-query and multi-query settings in Table 2. Our RMCA achieves comparable results with other methods. Although the rank-1 and mAP performance of RMCA under multi-query setting are slightly inferior to HA-CNN [17], it achieves the best rank-1 matching rate of 91.9% and mAP of 78.6% under single-query setting. Specifically, our model improves the single-query mAP by 2.0%, which indicates that the co-attention mechanism is not only capable of searching the most obvious ground truth match, but also is robust for diverse multi-camera pedestrian variations.

TABLE 2. Comparison of state-of-the-art methods on Market-1501 with both single-query and multi-query settings. The cumulative matching scores (%) at rank-1 and mAP (%) are listed.

Method	Single-Query		Multi-Query	
	Rank=1	mAP	Rank=1	mAP
SDALF [4]	20.5	8.2	29.2	13.8
eSDC [49]	33.5	13.5	42.5	18.4
LOMO+XQDA [1]	43.8	22.2	54.1	28.4
TMA [50]	47.9	22.3	–	–
DNS [51]	61.1	35.7	71.6	46.0
S-LSTM [23]	–	–	61.6	35.3
Gated-SCNN [36]	65.9	39.7	76.0	48.5
MSCAN [52]	80.3	57.5	86.8	66.7
SVDNet [53]	82.3	62.1	–	–
DPFL [31]	88.9	73.1	92.3	80.7
CAN [18]	60.3	35.9	72.1	47.9
IDEAL [40]	86.7	67.5	91.3	76.2
HA-CNN [20]	91.2	75.7	**93.8**	**82.8**
DuATM [17]	91.4	76.6	–	–
RMCA	**91.9**	**78.6**	93.6	82.2

To validate the statistical significance of our model performance, we execute a Wilcoxon signed-rank test on the Market-1501 results. The test verifies that the improvements in accuracy and mAP rates are statistically significant at the 4% significance level.

2) EXPERIMENTS ON CUHK03

As shown in Table 1, the CUHK03 dataset [5] contains 13,164 images of 1,360 pedestrians captured from six non-overlapping cameras over months. Each person is observed by two disjoint camera views and with approximately 2∼5 images in each view. This dataset is constructed by both manually labeled pedestrians and automatically detected bounding boxes. Examples of CUHK03 dataset are demonstrated in Fig. 3(b) and Fig. 3(c). The automatically detected dataset has a realistic setting because misalignment, occlusions, and missing body parts are common in this dataset. We follow the standard protocol [5]. That is, the dataset is partitioned into 1,260 persons for training and 100 persons for test. The experiments are conducted with 20 random splits for calculating the average performance. The CMC curve is

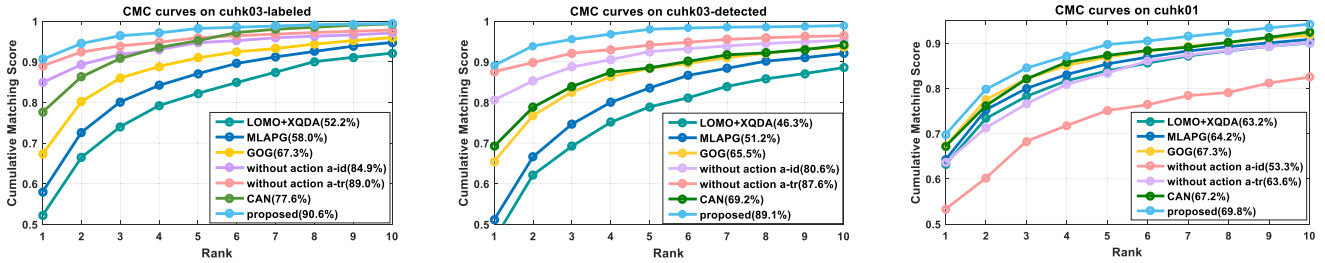


FIGURE 4. CMC curves and rank-1 identification rates for methods comparison on (a) the CUHK03 dataset with manually labeled pedestrian bounding boxes, (b) the CUHK03 dataset with automatically detected bounding boxes and (c) the CUHK01 dataset with 485/486 split.

TABLE 3. Comparison of state-of-the-art methods on CUHK03 with labeled setting. The cumulative matching scores (%) at rank-1, 5, 10 and 20 are listed.

Method	Rank1	Rank5	Rank10	Rank20
ITML [54]	5.5	18.9	30.0	44.2
SDALF [4]	5.6	23.5	36.1	52.0
LMNN [14]	7.3	21.0	32.0	48.9
eSDC [49]	8.8	24.1	38.3	53.4
KISSME [55]	14.2	48.5	52.6	70.0
FPNN [5]	20.7	51.5	68.7	83.1
IDML [6]	54.7	86.5	91.5	97.3
LOMO+XQDA [1]	52.2	82.2	92.1	96.3
MLAPG [33]	58.0	87.1	94.7	–
Ensembles [56]	62.1	89.1	94.3	97.8
DNS [51]	62.6	90.1	94.8	98.1
GOG [2]	67.3	91.0	96.0	98.8
MSE-VCM [57]	71.1	91.8	96.1	–
FT-JSTL+DGD [58]	75.3	–	–	–
DPFL[31]		–	–	–
CAN [18]	77.6	95.2	99.3	**100.0**
RMCA	**90.6**	**98.6**	**99.4**	99.8

TABLE 4. Comparison of state-of-the-art methods on CUHK03 with detected setting. The cumulative matching scores (%) at rank-1, 5, 10 and 20 are listed.

Method	Rank1	Rank5	Rank10	Rank20
SDALF [4]	4.9	21.2	35.1	48.4
ITML [54]	5.1	17.9	28.3	43.1
LMNN [14]	6.3	18.7	29.1	45.0
eSDC [49]	7.7	21.9	35.0	50.1
KISSME [55]	11.7	31.2	49.0	65.6
FPNN [5]	19.9	50.0	64.0	78.5
IDML [6]	45.0	76.0	83.5	93.2
LOMO+XQDA [1]	46.3	78.9	88.6	94.3
MLAPG [33]	51.2	83.6	92.1	96.9
DNS [51]	54.7	86.8	94.8	95.2
S-LSTM [23]	57.3	80.1	88.3	–
GOG [2]	65.5	88.4	93.7	97.6
Gated-SCNN [36]	61.8	80.9	88.3	–
MSE-VCM [57]	68.3	90.1	94.1	–
DPFL[31]	82.0	–	–	–
CRAFT-MFA [59]	84.3	97.1	98.3	99.1
CAN [18]	69.2	88.5	94.1	97.8
IDEAL [40]	71.0	89.8	93.0	95.9
RMCA	**89.1**	**98.0**	**98.9**	**99.6**

computed with the single-shot setting for both labeled and detected dataset.

According to Fig. 4(a) and Table 3, the proposed method achieves the best rank-1, rank-5 and rank-10 recognition rate of 90.6%, 98.6% and 99.4% on labeled CUHK03 dataset, although the rank-20 matching rate is slightly inferior to CAN [18]. Fig. 4(d) and Table 4 present the performances of the proposed method and other state-of-the-art methods using automatically detected bounding boxes. The result of the detected CUHK03 is usually inferior to the labeled CUHK03 due to the misalignment and occlusions caused by the detector, and the performances of other methods drop significantly, such as another attention based method CAN [18] dropping in rank-1 identification rate by 8.4%. However, our RMCA only decreases by 1.5% and surpasses the second best CRAFT-MFA [59] by 4.8%, which exhibits that RMCA is more robust to misalignment and occlusions.

3) EXPERIMENTS ON CUHK01

We further evaluate RMCA on the CUHK01 dataset [30]. As shown in Table 1, this dataset only contains 971 persons collected from two camera views in a campus environment. Each person only has two images in each camera view. The images in camera A are usually the side views of pedestrians, while the images in camera B are nearly the frontal or back view of pedestrians. It is obvious that the population size of CUHK01 is much smaller than Market-1501 and CUHK03 datasets. In addition, we follow a challenging split protocol [1], [18], [62] containing 485 persons for the training and 486 for the test. The random training/test split procedure is repeated 10 times and the average of CMC performance in multi-shot test setting [1] is reported in Fig. 4(c) and Table 5. CUHK01 is challenging for our method, since the small number of samples cannot allow to mine the full potential of RL. Typically, RL does need large data for model training. The evaluation on CUHK01 is a "potential/extreme"test, which however is not given in most existing re-id attention methods, i.e. our test is more comprehensive in this sense. And some hand-crafted re-id methods are very competitive

TABLE 5. Comparison of the state-of-the-art methods on CUHK01 with 486 test identities. The cumulative matching scores (%) at rank-1, 5, 10 and 20 are listed.

Method	Rank1	Rank5	Rank10	Rank20
LMNN [14]	13.5	31.3	42.3	54.1
ITML [54]	16.0	35.2	45.6	59.8
KISSME [55]	16.4	–	51.5	64.3
eSDC [49]	19.7	32.7	40.3	50.6
LFDA [60]	22.1	41.6	53.9	64.5
Mid-level Filter [61]	34.3	55.1	65.0	75.0
IDML [6]	47.5		80.0	–
LOMO+XQDA [1]	63.2	83.9	90.0	94.2
MLAPG [33]	64.2	85.5	90.8	94.9
multi-HG [3]	64.4	–	90.6	94.6
FT-JSTL+DGD [58]	66.6	–	–	–
GOG[2]	67.3	86.9	91.8	95.9
DNS [51]	69.1	86.9	91.8	95.4
MSE-VCM [57]	**70.5**	89.1	94.0	97.2
CAN [18]	67.2	87.3	92.5	97.2
RMCA	69.8	**89.7**	**94.2**	**97.4**

with their respective strengths particularly on small datasets like CUHK01. Encouragingly, our proposed method behaves robustly and outperforms most competitors except for the rank-1 of MSE-VCM [57].

4) COMPARISON TO EXISTING ATTENTION METHODS

To further demonstrate the effectiveness of our co-attention mechanism based on reinforcement learning, we specifically make a comparison between our method and several existing attention methods in the Table 2, Table 3, Table 4 and Table 5, including CAN [18], HA-CNN [20], DuATM [17] and IDEAL [40]. CAN [18] generates the attention regions by masking CNN feature map. In DuATM [17], there is no correlation across the temporal attention regions. HA-CNN [20] jointly learns the soft pixel attention and hard regional attention. IDEAL [40] just use RL to post-detect auto-detected bounding boxes to assist re-id. The existing attention methods are not quite consistent with human vision perception. Instead, we first introduce the temporal co-attention mechanism to simulate the scan order of human eyes and adopt reinforcement learning to perform this process. Since reinforcement learning provides a flexible learning strategy for sequential decision-making. The results also suggest the superiority of our RMCA over other attention methods.

C. EXPERIMENTAL ANALYSIS OF THE PROPOSED METHOD

1) EFFECT OF TIME STEP LENGTH

We evaluate the rank-1 identification rates using different number of time steps on both detected and labeled CUHK03 dataset, and the step number varies from 2 to 7. As shown in Fig. 5, the identification rates are gradually improved when the number of time steps increases from

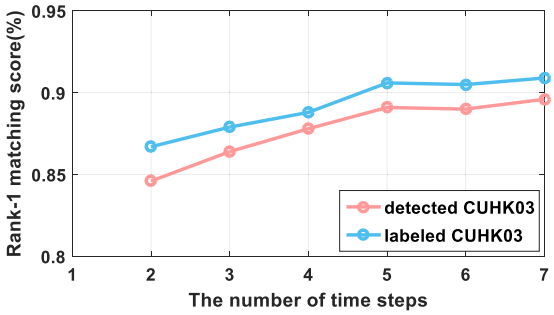


FIGURE 5. The rank-1 identification rate of our proposed method on labeled CUHK03 and detected CUHK03 datasets. The number of time steps varies from 2 to 7.

2 to 5, while more steps bring no obvious improvement instead of extra computation and complexity. Therefore, it is a tradeoff to choose 5 steps in our model for the accuracy and computational cost.

2) EVALUATION OF ACTION DESIGN

At the last step T of the sequential decision process, we introduce the identification action a_{id} and the triplet ranking action a_{tr} to facilitate the agent to interact with the pairwise image environment and locate the optimal co-attention regions. The combination of both actions is complementary. We evaluate the performance of the proposed method without the action a_{id} or a_{tr} to analyze their importance in our method. We also conduct the experiments only adopting the hybrid loss rather than co-attention mechanism.

As shown in Table 6, the accuracies of the comparative results would decline without the identification action a_{id}, the triplet ranking action a_{tr}, or the co-attention mechanism. And the lack of a_{id} would lead to further decline compared with the absence of a_{tr}. All these results indicate that the combined actions of a_{id} and a_{tr}, as well as co-attention mechanism are helpful for improving the accuracy. The identification action a_{id} makes better use of the multi-shot character of three datasets (each identity has multiple images), and is more about learning discriminative features from the image-identity correspondence relations. The triplet ranking action a_{tr} enables the agent to mine the hard positive and negative samples for learning critical information. The co-attention mechanism enables the agent to adaptively locate the discriminative regions.

3) VERIFICATION OF IMAGES WITH OCCLUSION

In the test phase, to observe the learned sequential co-attention processes, we randomly save some image pairs with co-attention region boxes. And we find that these region boxes are generally located in the salient parts of images and adaptively keep away from the occlusion, which is in accordance with the objective of reinforcement learning. Therefore, our co-attention mechanism is potentially robust to occlusions, which may be in a minority but are the most difficult cases in person re-id.

The following experiments are designed to evaluate the specific impact of occlusions to our RMCA and other

TABLE 6. Performance comparison (%) on Market-1501, CUHK03 and CUHK01, adopting different actions in reinforcement learning, or performing without co-attention mechanism. The cumulative matching scores (%) at rank-1, 5, 10 and 20, as well as mAP (%) are listed.

Method	Market-1501 (SQ)		Market-1501 (MQ)		CUHK03 (labeled)				CUHK03 (detected)				CUHK01 (485/486 split)			
	r=1	mAP	r=1	mAP	r=1	r=5	r=10	r=20	r=1	r=5	r=10	r=20	r=1	r=5	r=10	r=20
Without Action a_{id}	79.1	61.6	81.7	64.4	84.9	94.7	97.1	98.3	80.6	92.5	95.3	97.6	53.3	75.1	82.5	89.5
Without Action a_{tr}	89.6	74.8	90.8	77.5	89.0	95.9	97.8	98.9	87.6	94.1	96.4	97.1	63.6	83.4	90.1	94.8
Without co-attention	90.7	76.8	91.5	79.2	89.5	97.6	98.4	99.7	88.2	96.8	97.6	98.9	–	–	–	–
RMCA	**91.9**	**78.6**	**93.6**	**82.2**	**90.6**	**98.6**	**99.4**	**99.8**	**89.1**	**98.0**	**98.9**	**99.6**	**69.8**	**89.7**	**94.2**	**97.4**

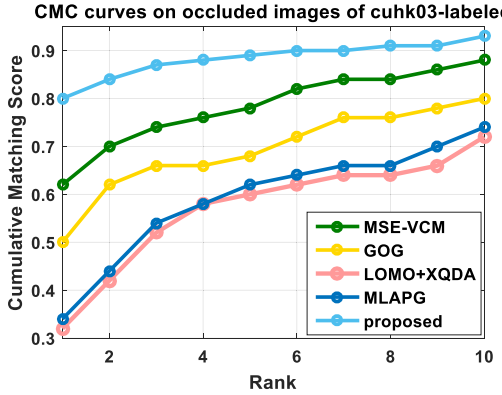


FIGURE 6. CMC curves for methods comparison on occluded images of labeled CUHK03 dataset(one split).

FIGURE 7. Visualizing the co-attention regions of different time steps learned by our method on four occlusion test samples from the labeled CUHK03 dataset. Each column indicates a time step, and for each orange frame, the left-most image pair is at the first time step.

four methods, including LOMO + XQDA [1], MLAPG [33], GOG [2] and MSE-VCM [57]. First, we choose one split from the experiments on labeled CUHK03 dataset. Then, we pick out 50 occluded images from its test probe images while remaining the same 100 test gallery images. Finally, the comparison experiments are conducted on the particular test dataset. As shown in Fig. 6, the rank-1 matching rates of LOMO + XQDA, MLAPG, GOG, GLM-VCM and RMCA are 32%, 34%, 50%, 62% and 79%, respectively. They are compared with those using all 100 probe images, decreasing 21%, 23%, 19%, 10% and 7%, respectively. Thus it can be seen that, compared to other four methods, our RMCA is less influenced by occlusions due to the co-attention mechanism based on reinforcement learning. In Fig. 7, we visualize the co-attention regions of different time steps learned by our method on three occlusion test samples from the labeled CUHK03 dataset. These co-attention processes support the conclusion that RMCA is able to ignore occlusions while concentrating on the relevant regions during eye movements.

V. CONCLUSION

In this paper, we propose to apply reinforcement learning based on RNN architecture to simulate local co-attention matching in person re-id. For any pedestrian image pair, the proposed model can adaptively select the optimal sequence of co-attention regions through the agent that interacts with the pairwise image environment. Extensive experiments on three challenging datasets demonstrate that our method achieves the comparable performance versus the state-of-the-art methods, especially increases robustness to occlusion cases. In the future, we will augment the model with another action. The added action would allow the agent to focus on the local regions of different scales toward various image pairs, which will be more flexible and targeted for matching image pairs.

REFERENCES

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

[2] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato, "Hierarchical Gaussian descriptor for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1363–1372.

[3] L. An, X. Chen, S. Yang, and X. Li, "Person re-identification by multi-hypergraph fusion," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 11, pp. 2763–2774, Nov. 2017.

[4] L. Bazzani, M. Cristani, and V. Murino, "Symmetry-driven accumulation of local features for human characterization and re-identification," *Comput. Vis. Image Understand.*, vol. 117, no. 2, pp. 130–144, 2013.

[5] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep filter pairing neural network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 152–159.

[6] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3908–3916.

[7] H. Wang, X. Zhu, S. Gong, and T. Xiang, "Person re-identification in identity regression space," *Int. J. Comput. Vis.*, vol. 126, no. 12, pp. 1288–1310, 2018.

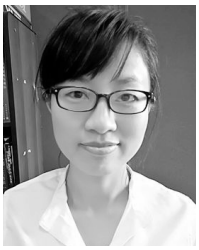
[8] A. Wu, W.-S. Zheng, and J.-H. Lai, "Robust depth-based person re-identification," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2588–2603, Jun. 2017.

[9] X. Li, M. Ye, Y. Liu, and C. Zhu, "Adaptive deep convolutional neural networks for scene-specific object detection," *IEEE Trans. Circuits Syst. Video Technol.*, to be published, doi: 10.1109/TCSVT.2017.2749620.

[10] X. Li, M. Ye, Y. Liu, F. Zhang, D. Liu, and S. Tang, "Accurate object detection using memory-based models in surveillance scenes," *Pattern Recognit.*, vol. 67, pp. 73–84, Jul. 2017.

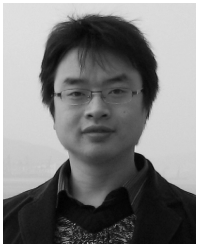
[11] Z. Zhang, T. Si, and S. Liu, "Integration convolutional neural network for person re-identification in camera networks," *IEEE Access*, vol. 6, pp. 36887–36896, 2018.
[12] J. Li, K. Lu, Z. Huang, L. Zhu, and H. T. Shen, "Transfer independently together: A generalized framework for domain adaptation," *IEEE Trans. Cybern.*, to be published, doi: 10.1109/TCYB.2018.2820174.
[13] J. Li, Y. Wu, and K. Lu, "Structured domain adaptation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 8, pp. 1700–1713, Aug. 2017.
[14] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *J. Mach. Learn. Res.*, vol. 10, pp. 207–244, Feb. 2009.
[15] X. Li, L. Liu, and X. Lu, "Person reidentification based on elastic projections," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 4, pp. 1314–1327, Apr. 2018.
[16] C. Zhang and Q. Liu, "Region constraint person re-identification via partial least square on Riemannian manifold," *IEEE Access*, vol. 6, pp. 17060–17066, 2018.
[17] J. Si *et al.*, "Dual attention matching network for context-aware feature sequence based person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5363–5372.
[18] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan, "End-to-end comparative attention networks for person re-identification," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3492–3506, Jul. 2017.
[19] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person Re-identification by discriminative selection in video ranking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 12, pp. 2501–2514, Dec. 2016.
[20] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2285–2294.
[21] 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 *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 4743–4752.
[22] R. J. Williams and D. Zipser, "A learning algorithm for continually running fully recurrent neural networks," *Neural Comput.*, vol. 1, no. 2, pp. 270–280, 1989.
[23] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang, "A siamese long short-term memory architecture for human re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 135–153.
[24] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, p. 529, 2015.
[25] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey," *J. Artif. Intell. Res.*, vol. 4, no. 1, pp. 237–285, Jan. 1996.
[26] R. Zhao, W. Ouyang, and X. Wang, "Person re-identification by saliency learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 2, pp. 356–370, Feb. 2017.
[27] V. Mnih, N. Heess, and A. Graves, "Recurrent models of visual attention," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 2204–2212.
[28] M. Hayhoe and D. Ballard, "Eye movements in natural behavior," *Trends Cognit. Sci.*, vol. 9, no. 4, pp. 188–194, 2005.
[29] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1116–1124.
[30] W. Li, R. Zhao, and X. Wang, "Human reidentification with transferred metric learning," in *Proc. Asian Conf. Comput. Vis.*, 2012, pp. 31–44.
[31] Y. Chen, X. Zhu, and S. Gong, "Person re-identification by deep learning multi-scale representations," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop*, Oct. 2018, pp. 2590–2600.
[32] J. Jiao, W.-S. Zheng, A. Wu, X. Zhu, and S. Gong, "Deep low-resolution person re-identification," in *Proc. AAAI*, 2018, pp. 6967–6974.
[33] S. Liao and S. Z. Li, "Efficient PSD constrained asymmetric metric learning for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3685–3693.
[34] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[35] A. Hermans, L. Beyer, and B. Leibe. (2017). "In defense of the triplet loss for person re-identification." [Online]. Available: https://arxiv.org/abs/1703.07737
[36] R. R. Varior, M. Haloi, and G. Wang, "Gated siamese convolutional neural network architecture for human re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 791–808.
[37] S. Mathe, A. Pirinen, and C. Sminchisescu, "Reinforcement learning for visual object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2894–2902.
[38] J. C. Caicedo and S. Lazebnik, "Active object localization with deep reinforcement learning," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 2488–2496.
[39] M. Malmir, K. Sikka, D. Forster, I. Fasel, J. R. Movellan, and G. W. Cottrell, "Deep active object recognition by joint label and action prediction," *Comput. Vis. Image Understand.*, vol. 156, pp. 128–137, Mar. 2017.
[40] X. Lan, H. Wang, S. Gong, and X. Zhu. (2017). "Deep reinforcement learning attention selection for person re-identification." [Online]. Available: https://arxiv.org/abs/1707.02785
[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.
[42] Q. Yu, X. Chang, Y.-Z. Song, T. Xiang, and T. M. Hospedales. (2017). "The devil is in the middle: Exploiting mid-level representations for cross-domain instance matching." [Online]. Available: https://arxiv.org/abs/1711.08106
[43] D. Wierstra, A. Foerster, J. Peters, and J. Schmidhuber, "Solving deep memory pomdps with recurrent policy gradients," in *Proc. Int. Conf. Artif. Neural Netw.*, 2007, pp. 697–706.
[44] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Mach. Learn.*, vol. 8, nos. 3–4, pp. 229–256, 1992.
[45] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2000, pp. 1057–1063.
[46] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2818–2826.
[47] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[48] N. Mclaughlin, J. M. del Rincon, and P. Miller, "Recurrent convolutional network for video-based person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1325–1334.
[49] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3586–3593.
[50] N. Martinel, A. Das, C. Micheloni, and A. K. Roy-Chowdhury, "Temporal model adaptation for person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 858–877.
[51] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1239–1248.
[52] D. Li, X. Chen, Z. Zhang, and K. Huang, "Learning deep context-aware features over body and latent parts for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 384–393.
[53] Y. Sun, L. Zheng, W. Deng, and S. Wang, "SVDNet for pedestrian retrieval," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 3820–3828.
[54] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *Proc. Int. Conf. Mach. Learn.*, 2007, pp. 209–216.
[55] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2288–2295.
[56] S. Paisitkriangkrai, C. Shen, and A. van den Hengel, "Learning to rank in person re-identification with metric ensembles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1846–1855.
[57] L. Lin, R. Huang, X. Li, F. Zhang, and M. Ye, "Person re-identification by optimally organizing multiple similarity measures," *IEEE Access*, vol. 5, pp. 26034–26045, 2017.
[58] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1249–1258.
[59] Y.-C. Chen, X. Zhu, W.-S. Zheng, and J.-H. Lai, "Person re-identification by camera correlation aware feature augmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 2, pp. 392–408, Feb. 2018.
[60] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian, "Local Fisher discriminant analysis for pedestrian re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3318–3325.

[61] R. Zhao, W. Ouyang, and X. Wang, "Learning mid-level filters for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 144–151.
[62] W. Li, X. Zhu, and S. Gong, "Person re-identification by deep joint learning of multi-loss classification," in *Proc. Int. Joint Conf. Artif. Intell.*, 2017, pp. 2194–2200.



LAN LIN received the B.S. and M.S. degrees from the University of Electronic Science and Technology of China, Chengdu, China, in 2006 and 2010, respectively, where she is currently pursuing the Ph.D. degree. Her current research interests include machine learning and computer vision.

HUAN LUO received the B.S. degree in mechanical and electronic engineering from Northwestern Polytechnical University, in 2016. He is currently pursuing the master's degree with the University of Electronic Science and Technology of China, Chengdu, China. His research interests include pattern recognition and machine learning.



RENJIE HUANG received the master's degree in computer science from Southwest University, in China, in 2006, and the Ph.D. degree from the University of Electronic Science and Technology of China, Chengdu, China, in 2015. He is currently an Instructor with the School of Computer and Information Science, Southwest University, China. His research interests include machine learning and computer vision.

MAO YE received the B.S. degree from Sichuan Normal University, Chengdu, China, in 1995, the M.S. degree from the University of Electronic Science and Technology of China, Chengdu, China, in 1998, and the Ph.D. degree from The Chinese University of Hong Kong, China, in 2002, all in mathematics. He has been a short-time Visiting Scholar with The University of Queensland, and the University of Pennsylvania. He is currently a Professor and the Director of CVLab, University of Electronic Science and Technology of China. His research interests include machine learning and computer vision. In these areas, he has published over 90 papers in leading international journals or conference proceedings. He was a co-recipient of the Best Student Paper Award from the IEEE ICME 2017. He has served on the Editorial Board of the *Engineering Applications of Artificial Intelligence*.