Context-Interactive CNN for Person Re-Identification

Wenfeng Song, Shuai Li, Tao Chang, Aimin Hao, Qinping Zhao, and Hong Qin

***Abstract*— Despite growing progresses in recent years, cross-scenario person re-identification remains challenging, mainly due to the pedestrians commonly surrounded by highly-complex environment contexts. In reality, the human perception mechanism could adaptively find proper contextualized spatial-temporal clues towards pedestrian recognition. However, conventional methods fall short in adaptively leveraging the long-term spatial-temporal information due to ever-increasing computational cost. Moreover, CNN-based deep learning methods are hard to conduct optimization due to the non-differentiable property of the built-in context search operation. To ameliorate, this paper proposes a novel Context-Interactive CNN (CI-CNN) to dynamically find both spatial and temporal contexts by embedding multi-task Reinforcement Learning (MTRL). The CI-CNN streamlines the multi-task reinforcement learning by using an actor-critic agent to capture the temporal-spatial context simultaneously, which comprises a context-policy network and a context-critic network. The former network learns policies to determine the optimal spatial context region and temporal sequence range. Based on the inferred temporal-spatial cues, the latter one focuses on the identification task and provides feedback for the policy network. Thus, CI-CNN can simultaneously zoom in/out the perception field in spatial and temporal domain for the context interaction with the environment. By fostering the collaborative interaction between the person and context, our method could achieve outstanding performance on various public benchmarks, which confirms the rationality of our hypothesis, and verifies the effectiveness of our CI-CNN framework.**

***Index Terms*— Person re-identification, multi-task reinforcement learning, context interaction, actor-critic agent, context-critic network.**

Manuscript received January 30, 2019; revised October 6, 2019; accepted November 8, 2019. Date of publication November 20, 2019; date of current version January 23, 2020. This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB1700603, in part by the National Natural Science Foundation of China under Grant 61672077 and Grant 61532002, in part by the Applied Basic Research Program of Qingdao under Grant 161013xx, in part by the Beijing Natural Science Foundation-Haidian Primitive Innovation Joint Fund under Grant L182016, in part by the National Science Foundation of USA under Grant IIS-0949467, Grant IIS-1047715, Grant IIS-1715985, and Grant IIS-1049448, and in part by the Capital Health Research and Development of Special under Grant 2016-1-4011. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Guo-Jun Qi. *(Corresponding authors: Shuai Li; Hong Qin.)*

W. Song, T. Chang, A. Hao, and Q. Zhao are with the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China.

S. Li is with the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China, and also with the Peng Cheng Laboratory, Shenzhen 518052, China (e-mail: lishuai@buaa.edu.cn).

H. Qin is with the Department of Computer Science, Stony Brook University, Stony Brook, NY 11794-4400 USA (e-mail: qin@cs.stonybrook.edu).

This article has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors.

Digital Object Identifier 10.1109/TIP.2019.2953587

I. Introduction

PERSON re-identification (Re-ID) gains growing momentum recently, which is fundamental for environment monitoring, search/rescue, smart surveillance, and some wearable devices based applications. In particular, cross-scenario Re-ID aims to automatically match pedestrians captured by cameras at different locations or time, which requires the identification model to be resourceful over different target datasets. Thus, cross-scenario Re-ID still has many challenges yet to overcome. The pivotal challenge is how to capture the cross-scenario context information around the designated pedestrian. Specially, the drastically-changing camera views, clutter background, low resolution, and other objects occlusion tend to cause ambiguity for the identification.

Existing Re-ID methods typically focus on suppressing the background effects in the spatial domain [1], [2]. Such methods usually resort to separate handling of the background and foreground in single image. Their key idea is to find the person related regions that are coherent across different scenarios. However, in practice, it is hard to obtain satisfactory performance by suppressing the clutter background, because the blurry motion, low resolution [3], [4] and heavy occlusion in the unconstrained real-world scenarios would destroy the pedestrian's integrity when extracting discriminative features. In summary, single image/pedestrian's region is insufficient for the cross-scenario Re-ID task, too much or too less context can influence the feature extraction, only proper context could facilitate to improve the performance. Besides, the proper use of background [5] could also improve the final recognition result. Intuitively, the spatial context indicates the relations between the surrounding background and the target pedestrian. For example, the pedestrians walking on the road and the persons riding on the bikes are discriminately in two main pose shapes. Secondly, the buildings and trees often partially occlude the pedestrians. Thirdly, the accessories, like bags and umbrellas, may provide auxiliary clues to determine the pedestrian's features.

On the other hand, the temporal context clues could also improve the performance. To efficiently leverage the temporal information, recently, LSTM or RNN based methods [6]–[9] have been proposed to aggregate the temporal context features in the videos. However, such kind of methods commonly treat all the frames with equal weights, which results in high computational cost and slow convergence for different-structure networks. Thus, some works [10]–[12] propose to pay more attention to the key frames. Nevertheless, these

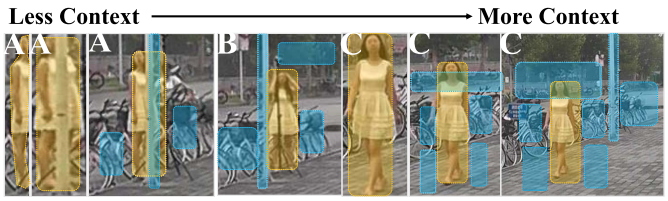


Fig. 1. Illustration of the temporal-spatial context. For an identical Re-ID model, we use different bounding boxes related to objects (shown in yellow) and context (shown in blue) to provide more coherent cues than naively removing the non-person pixels. A, B, C denote different frames of same pedestrian.

methods cannot make a good balance between the background and person's foreground region. Recently, the most relevant works [13], [14] similar to our approach have already achieved satisfying performance by modeling the temporal and spatial relevant motions of individuals while suppressing the irrelevant motions in human interaction and activity recognition. These works motivate us to improve the method for the person Re-ID task. Moreover, we observe that, aggregating pedestrian's region in temporal sequences would perform better than leveraging single image information, because single image is ambiguous to delimit irrelevant regions, while temporal sequences give rise to better suppressing results for irrelevant backgrounds.

Despite existing works investigate the influence of the background, they rarely study how to dynamically find the proper spatial-temporal context for each specific pedestrian region. Since the human vision and memory systems recognize persons by comprehensively judging consecutive actions and surrounding environment context in a temporal-spatial collaborating way, to mimic this mechanism, we design a novel context-aware model to take advantages of the temporal-spatial context information. Our key insight is that, given discriminative pedestrian and relevant context in the vicinity, Re-ID model should be adaptive to different scenarios. The appearance features in RGB image will change dramatically across scenarios. However, the context relations of the pedestrians are stable. Meanwhile, to fit for the proper range of context, the fixed perception field in the convolution layers are dynamically zoomed-in/out by the context's wrapping operator, as shown in Fig. 2. Then the network could dynamically extract the contexts features of the current images based on the discrimination ability. Moreover, the temporal context can provide intrinsic clues for the occluded and pose-variant pedestrians. We embed the temporal clues by decomposing the frames into background and pedestrian-relevant components in proper temporal context ranges. As for the discriminative representation of pedestrian, it has been well solved by CNN methods [15]. We mainly focus on exploiting the contextual cues in this paper. We use an adaptive spatial-temporal context learning process to provide more coherent clues, rather than naively removing the non-person pixels.

In some sense, the Re-ID task on different persons can be considered as the interaction between persons and the spatial-temporal context, wherein the pedestrian's regions provide appearance information, and the surrounding regions and relevant temporal sequences provide the optimal context

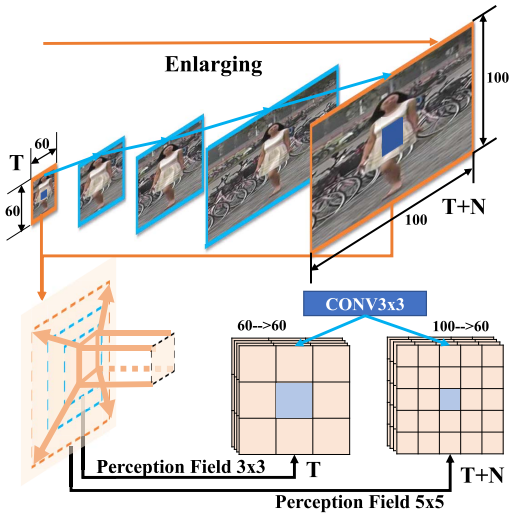


Fig. 2. The perception field changes with the context of pedestrian's images. The image with context is enlarged from 60*60 to 100*100. When filtered with the 3*3 convolution, the perception field is enlarged from 3*3 to 5*5.

priors to distinguish the foreground object from its background (Fig. 1). Different from most existing methods, which focus on spatial policy while ignoring temporal clues or vice versa, we propose a multi-task reinforcement learning framework, which jointly learns the spatial-temporal clues. Given an single image, we can determine whether we need a larger region or longer video sequences. Since the two tasks share the same states in reinforcement learning, we can combine the two tasks into an single actor-critic agent. This agent takes actions (i.e., *enlarge, shrink, or stay the same*) to determine the size of the pedestrian's bounding box and the length of the consecutive frames to provide proper context ranges for pedestrian identification. Specially, the salient contributions of this paper can be summarized as follows.

- We propose to solve the cross-scenario person Re-ID problem by integrating novel actor-critic agent in reinforcement learning embedded CNN, which can dynamically find the contexts and telescoping perception field, and thus gives rise to intrinsic information for person identification.
- We reveal the relations between the context and pedestrian for cross-scenario Re-ID, and formulate the interaction between optimal person regions and their spatial-temporal context, which gives rise to simple yet effective learning policies for obtaining optimal context clues.
- We design a novel deep reinforcement learning based multi-task framework to learn the context-agent by collaborating with the context policy network and context-critic network. Comprehensive experiments have demonstrated the superiorities of our framework.

II. Related Works

A. Person Re-ID

Recently, deep learning based methods [16]–[18] have been dominating the person Re-ID research community, benefiting from their superior discriminative ability. Image-based methods usually emphasize to improve the feature representation

that is insensitive to pose, clutter background, and camera views. For example, Liu et al. and Liang eta al. proposed to extracte pose-invariant [16], [19]–[21] features via alignment of pedestrians and suppressing the background noise. Meanwhile, Zheng *et al.* [22] proposed camera view invariant model. Feng *et al.* [23] and *Borgia et al.* [24] proposed view-specific deep networks. To further extract the scenarios-invariant features, the spatial based attention models [1], [2], [17], [25], [26] are proposed to evaluate the importance of different regions and its surrounding context for pedestrian's feature's representation. Recently, in the work [5], the authors investigated the background's effect on the feature representation in Re-ID task. Besides, some works [27], [28] tried to search the query-related pedestrian in spatial context of the whole image. These works verify that different regions and their spatial context have different weights in Re-ID task. The previous works achieve satisfying performance for some specific scenarios. Nevertheless, image-based methods ignore the temporal relations, as a result, such methods lack the generalization ability to be further applied to un-known datasets.

To solve the cross-scenario problems, some video-based works [6], [29] try to capture temporal clues by exploiting the coherency from the videos. Recently, temporal and spatial combined attention based methods [10], [30] achieve impressive performance. Besides, self paced weighting [11] exploited different weights of frames to improve the Re-ID in condition of object occlusion and motions. However, these methods are limited in the single scenario and remains hard to capture sufficient context clues across scenarios. Recently, Tang *et al.* [13] proposed a novel Coherence Constrained Graph LSTM (CCG-LSTM) method to model the relevant motions of individuals while suppressing the irrelevant motions, which achieved state-of-the-art performance in recognizing the group activity. Likewise, Shu *et al.* [14] also proposed a novel Hierarchical Long Short-Term Concurrent Memory (H-LSTCM) method to model the long-term inter-related dynamics among a group of persons for recognizing human interactions. In terms with cross-scenario Re-ID task, the most relevant work with ours is unsupervised cross-dataset transfer learning for person Re-ID [31], [32]. Inspired by these works, we propose to capture both spatial and temporal contexts in an adaptive framework for each frame.

B. Reinforcement Learning

Reinforcement learning mainly attempts to solve the policy learning problems in some vision tasks, and achieve remarkable success in video parsing applications, including semantic segmentation [33], video caption [34], interaction cropping [35], and video summarization [36]. Zhong *et al.* [37] proposed hierarchical tracking by using reinforcement learning based searching and coarse-to-fine verifying, and achieved satisfying performance. Recently, Yang *et al.* [38] proposed to embed the multi-task into an single policy network to efficiently control the multiple actions. Inspired by the previous works, we try to adaptively find context cues for single image based person Re-ID task via Asynchronous Advantage Actor-Critic (A3C) [39], wherein we use a multi-task neural network to represent the actor-critic agent. To the best of our knowledge, we are among the first to employ reinforcement learning to deal with the person Re-ID problem by simultaneously learning the spatial and temporal contexts.

III. New Methodology and Network

In this section, we start with an overview of our CI-CNN, and then describe the details of the proposed multi-task reinforcement learning module and the carefully designed context-interactive network, after that, we detail the optimization, followed by the inference process.

A. Overview of Our CI-CNN

The pipeline of our CI-CNN is shown in Fig. 3, which employs an actor-critic agent to search the Re-ID task-oriented temporal and spatial context in the current pedestrian image. It involves two main components: 1) multi-task reinforcement learning (MTRL) agent consists of context-policy network $\pi(s_t;\theta)$ and context-critic network $V(s_t;\theta_v)$. Given image I serving as an element in state s_t, $\pi(s_t;\theta)$ takes in single image I serve as an element in state s_t and outputs the actions: a_t^S in spatial domain, a_t^T in temporal domain. Meanwhile, $V(s_t;\theta_v)$ provides corresponding temporal estimated error e_t to $\pi(s_t;\theta)$, which indicates the quality of the actions ('estimated error' denotes the error differences between two iterations of the context-critic network); 2) context-interactive network (CIN) $e(s_t;\theta_e)$ is designed to sufficiently utilize the spatial-temporal clues inferred from the actor-critic agent. In Fig. 3, we use boxes with different colors to distinguish the spatial (green) and temporal (blue) contexts.

B. Multi-Task Reinforcement Learning

We formulate the adaptive context search as an actor-critic process. In the search process, the actor-critic agent interacts with the environment, and takes a series of actions a_t^T and a_t^S to optimize both the spatial context and temporal context search policies via reward R from the environment. The two context search tasks are trained in an end-to-end way in single policy network. The flow chart of learning process is illustrated in Fig. 4. The agent first receives observations from the single image I, and then it samples action from the action space according to the state of MTRL which contains observations and historical experience. Thereafter, it executes the sampled action to determine the size of spatial context box and the length of the temporal sequences. After executing each action, it receives a reward according to the recognition accuracy of the newly generated images and sequences. By maximizing the accumulated rewards, the agent aims to find the most satisfied context sequences relevant with the current frame. In the following, we introduce the basic components of our actor-critic agent: environment, action space, state, and reward in a sequential order.

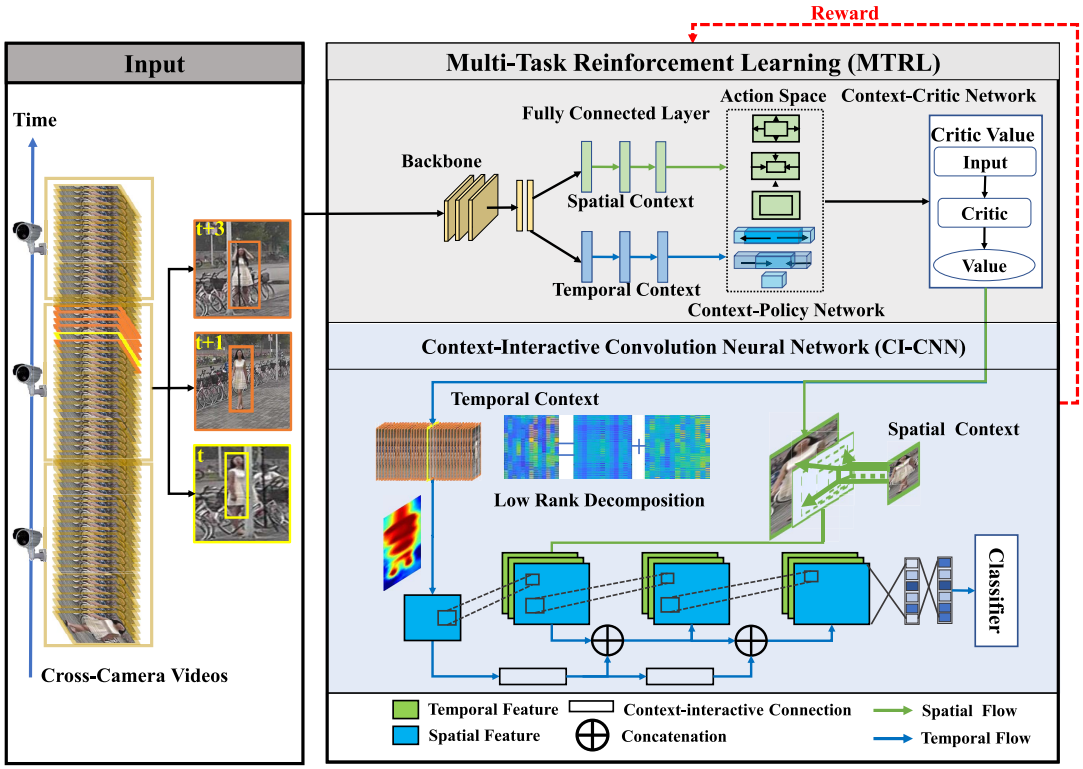


Fig. 3. Pipeline of our CI-CNN. MTRL: taking image as input, MTRL will output actions to adjust the size of the bounding box (i.e., *zoom-in* the pedestrian's box) and the length of neighboring frames (i.e., *enlarge* the consecutive frames). According to the actions, a new pedestrian box and a low-rank component in the suggested length of sequences will be computed. Person-Context Interaction: the context-interactive network (CIN) takes the image and its corresponding low-rank component as input, provides the reward to evaluate the actions, and updates the actor-critic agent.

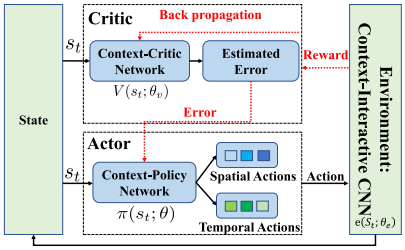


Fig. 4. Pipeline of MTRL network. The current state s_t is fed into context-policy network $\pi(s_t;\theta)$, it is split into two branches for two tasks: spatial and temporal search. The pedestrian with the new contexts are fed into the context-interactive network $e(s_t;\theta_e)$ to get reward from environment. Then $\pi(s_t;\theta)$ is updated from context-critic network $V(s_t;\theta_v)$ via estimated error, $V(s_t;\theta_v)$ is updated by the accumulated reward that is back propagated from $e(s_t;\theta_e)$.

1) Environment: The environment interacts with the proposed MTRL network via the pre-defined reward mechanism. Different from traditional reinforcement learning frameworks [39], [40], which maintain environment with fixed reward mechanism, our environment contains a context-interactive and dynamically-updated CNN $e(s_t;\theta_e)$ to provide the reward, wherein θ_e denotes the iteratively-updated parameters of CNN. In our training settings, the environment will be updated with the newly generated context of the agents.

2) Action Space: The multi-task action space $\mathcal{A}$ is supposed as corresponds to the set of actions which is affected by M search tasks. Mathematically, $\mathcal{A}$ is defined as,

$$\mathcal{A}(\cdot) = \prod_{j=1}^{M} \mathcal{A}_j. \tag{1}$$

Here $\prod$ denotes the Cartesian Product. In our experiments $M = 2$, we design 3 spatial actions $a_t^S \in \mathcal{A}_1 = \{$zoom in, zoom out, stay the same$\}$, 3 temporal context-changing actions $a_t^T \in \mathcal{A}_2 = \{$enlarge, shrink, stay the same$\}$. The total number of context actions is $3 * 3$. The termination action <stay the same, stay the same> will terminate the current episode, and then CIN will output the results immediately. Otherwise, the agent will continue to search more context clues until meeting the preset maximum time step t_{max}.

The reward function for the i-th task depends only on the corresponding actions. Therefore, each task-oriented action is independently trained in our network. In this way, the training processes of task-specific actions are decoupled. Essentially, this kind of parametrization is more efficient in the case of shared actions. Consequently, the number of parameters can be significantly reduced by the use of single network.

3) State and Reward: The state s_t of MTRL is represented as a tuple $s_t = (I_t, r_t, a_t)$, where $a_t = (a_t^S, a_t^T)$, the current RGB image is $I_t \in \mathbb{R}^{N\times N\times 3}$ (the selected frame in the t-th step of action), and the reward r_t is propagated from the context-interactive network. This tuple summarizes the history of observations and actions from the beginning of the search sequence: the pedestrian related spatial context

action history $\mathcal{H}_t^S$ records each a_t^S, the temporal context action history $\mathcal{H}_t^T$ records each a_t^T. These components are essential and complementary, because r_t indicates how accurate the recognition output is, while a_t provides visual cues for the agent to update the environment. The state s_t is formed by the context-critic network and updated over time by the context-policy network.

The context-critic network $V(s_t;\theta_v)$ gives the reward immediately based on the quality of recognition prediction after the agent chooses an action. We get the critic value from the context-critic network $V(s_t;\theta_v)$. The reward of the CI-CNN is defined as follows,

$$r_t = \begin{cases} \alpha \cdot sign[V(s_t;\theta_v) - V(s_{t+1};\theta_v)] & \text{if } a_t \neq \text{terminate}, \\ sign[\epsilon - V(s_t;\theta_v] & \text{otherwise.} \end{cases} \tag{2}$$

Here ϵ is a threshold to discriminate whether the improvement is sufficient or not, $sign[\cdot]$ returns 1 (-1) as positive (negative) value. The scale factor α and the threshold ϵ are set to 0.1 and 0.05 empirically. Eq. 2 indicates the agent receives a positive reward when the chosen action improves the recognition accuracy, and receives a penalty when it decreases the performance. If the agent chooses to terminate the process, the final recognition prediction must be good enough, otherwise it would receive large penalty.

C. Context Interaction

This section details the interaction process between the person and the context. Specially, based on the aforementioned four basic elements, we propose a multi-task actor-critic agent (as shown in Fig. 4) for context searching. We now depict the process to find the spatial and temporal context in single image, which is reduced to two tasks: searching the spatial context c_s around current pedestrian's box and searching the temporal context c_t in the neighboring frames. We integrate the two objectives into single actor-critic agent, which involves a multi-task context-policy network and a context-critic network.

1) Context-Policy Network: Given single image I, the context-policy network $\pi(s_t;\theta)$ should determine the two actions a_t^S and a_t^T. The previous work [38] demonstrates that single critic network is sufficient in multi-policy learning. Therefore, we simplify the multi-task $\pi(s_t;\theta)$ as a sharing backbone network with two side-by-side softmax layers after the fully-connected layer, the output actions are evaluated by the context-critic network. More concretely, the architecture of the context-policy network π is shown in the left part of Fig. 3. The network π uses a Vanilla densely-connected network [15] as the backbone (yellow parts), and outputs the operations for searching context ranges of related spatial regions and neighboring frames. Besides, the instability of the multi-task network is addressed by independently training the two separate cross-entropy losses.

2) Context-Critic Network: As for the context-critic network (as shown in Fig. 4), given the single image I and the actions predicted from the $\pi(s_t;\theta)$, the context-critic network $V(s_t;\theta_v)$ outputs the critic value to evaluate the actions. We propose to define the critic value via a convolutional neural network that has one softmax output for the context-critic $e(s_t;\theta_e)$ and one linear output for the critic value, with all layers shared. Therefore, the critic value represents the probability of the predicted person ID to be the true person ID. If the network increases probability with the context actions than that without the context, it will feedback positive reward, otherwise, it will feedback negative reward. Furthermore, the critic should evaluate both temporal and spatial actions. Towards this goal, we use the shared parameters in single critic framework, which could help the network capture the correlation among the actions much better than the multi-critic case. Note that, single critic framework can achieve these performances with fewer parameters than the multi-critic framework.

3) Optimization of MTRL: We propose to improve a variant of asynchronous advantage actor-critic (A3C) algorithm in [35] to train the MTRL. The conventional reinforcement learning methods generally adopt single task network architecture, instead, to make it satisfy our multi-task context-policy, the critic network outputs rewards of the two actions in parallel. We use r_t in Eq. 2 to denote the immediate reward at step t, and then the accumulated reward is $\sum_{i=0}^{k-1}\gamma^i r_{t+i} + \gamma^k V(s_{t+k};\theta_v)$, $V(s_t;\theta_v)$ is the output value under state s_t. Here θ_v denotes the network parameters of critic branch, and k ranges from 0 to t_{max} (maximum number of steps before updating). Therefore, the objective of the agent is to maximize the Expectation $\mathbb{E}$ of reward:

$$R(a_t, s_t) = \mathbb{E}[\frac{1}{N}\sum_{i=1}^{N}\sum_{t=0}^{\infty}\gamma^t r_t^i(s_t, \pi(s_t;\theta))]. \tag{3}$$

Here γ is a discount factor, which controls the effect of the state in long time, r_t is the immediate reward according to the current state s_t, N is the total actions number, and t denotes the t-th epoch.

Furthermore, the optimization objective of the policy output is to maximize the advantage function $R_t - V(s_t;\theta_v)$ and the entropy of the context-policy output $L(\pi(s_t;\theta))$. The cross entropy loss in the optimization objective is used to increase the diversity of actions, which can make the agent learn flexible policies. We use estimated error to compute the policy gradient. The gradients of the actor-policy $\pi(s_t;\theta)$ can be formulated as:

$$\nabla_\theta R = \nabla_\theta log\pi(a_t|s_t;\theta)((R - V(s_t;\theta_v)) + \beta\nabla_\theta L(\pi(s_t;\theta))). \tag{4}$$

Here the context-critic network $V(s_t;\theta_v)$ approximates the total expected reward R of state s_t with parameters θ_v. The optimization objective of the context-critic output is to minimize $(R - V(s_t;\theta_v))^2/2$. Thus, the gradients of the context-critic network $V(s_t;\theta_v)$ can be formulated as:

$$\nabla_{\theta_v} R = \nabla_{\theta_v}(R - V(s_t;\theta_v))^2, \tag{5}$$

where β controls the influence of entropy $L(\pi(s_t;\theta))$, and $\pi(s_t;\theta)$ is the probability of the sampled action a_t under the same state s_t.

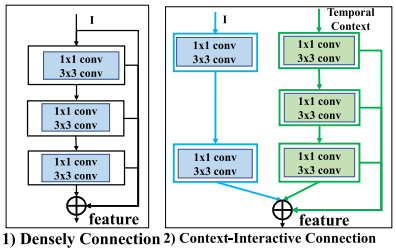


Fig. 5. The comparison between Vanilla dense connection and the context-interactive connection in our network, and $\bigoplus$ denotes the concatenation operation.

D. Context-Interactive Network for Person Re-ID

After we get the proper context of person regions, we exploit the efficient CNN to fully take advantage of the temporal and spatial contexts. We propose a novel context-interactive network scheme similar to Densenet. Different from it, we embed a low-rank component as an auxiliary channel to arouse the person related semantic features. We use the robust PCA [41] method to decompose the temporal context sequences into low-rank component I_l and sparse component I_s, which belong to $\mathbb{R}^{n\times n}$.

$$I = I_l + I_s, \tag{6}$$

which subjects to,

$$I_l = argmin\ ||I - I_l||_F, \quad s.t.\ rank(I_l) = r. \tag{7}$$

Here $||\ ||_F$ represents the Frobenius norm. r denotes the rank of sequences I, and we set it as 2. In this way, we can get the foreground component, which is the mean visual feature of the sequences, and the unrelated background can be removed from the sequences. We further employ I_l as the attention map for the image with the skipping connections, which are used to retain the context region information when activating deeper feature maps. The differences between the context-interactive connection and Vanilla dense connection are shown in Fig. 5. In our network, the new connection replaces the original one in the dense network, which demonstrates a better performance.

Our CI-CNN aims to find the most informative context ranges around the current frame. Inference is carried out by sampling from the policy $\pi(s_t, \theta)$ iteratively, until the termination action is taken. At each step, state s_t is updated according to action a_t. When the search is complete, the confidence of the spatial and temporal actions context is output by the context-policy network. As for inference, given a set of images together with the action confidence function $p(a_t^T|s_t, \theta)$, it should be maximal when identification is accurate. We want to find the parameters maximizing the Re-ID accuracy based on the context and confidence $p(a_t^T|s_t, \theta)$ in the last step. At the same time, we aim to minimize the number of the evaluated temporal frames. Collaborating with the context-critic network, the context-policy can progressively approach the optimal pedestrian box and context frames. By repeating this process, the proposed method can provide context for the whole video sequence. Some examples from the testing process are shown in Fig. 6.

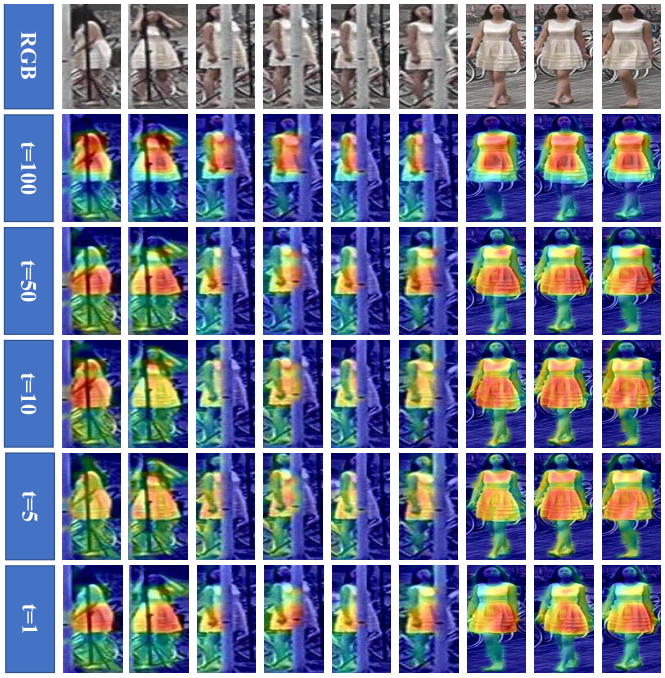


Fig. 6. Temporal context over time indicates the person-related temporal clues gradually focus on the pedestrian in spite of (possible) occluding regions. Heatmap follows HSV.

Algorithm 1 Training Procedure of Our CI-CNN

Input : Initialized cropped pedestrian images I and IDs
Output: CI-CNN model

Initialize S (for each episode) $e(I, s, \theta)$;
$I_0 \leftarrow I, t \leftarrow 0$
repeat
  A $\sim \pi(|S, \theta)$;
  Take action $a(s,t) \in \mathcal{A}$, observe S, R;
  $\delta \leftarrow R + \gamma\hat{v}(S, \theta_v) - \hat{v}(S, \theta_v)$;
  $z^{\theta_v} \leftarrow \gamma\lambda^{\theta_v} z^{\theta_v} + I\nabla_{\theta_v}\hat{v}(s, \theta_v)$;
  $z^{\theta} \leftarrow \gamma\lambda^{\theta} z^{\theta} + I\nabla_{\theta} ln\pi(a|S, \theta)$;
  $\theta_v \leftarrow \theta_v + \alpha^{\theta_v}\delta z^{\theta_v}$;
  $\theta \leftarrow \theta + \alpha^{\theta}\delta z^{\theta}$;
  $I \leftarrow \gamma I_{t-1}$;
  $S \leftarrow S_{t-1}$;
until $t - t_{start} == t_{max}$ *or* a_{t-1} *is termination*;
for $i \in I$ **do**
  $R \leftarrow r_i - \gamma R$;
  $d\theta \leftarrow d\theta + \nabla_{\theta} log\pi(a_i|s_i; \theta)(R - v(s_i; \theta_v))$;
  $+\beta\nabla_{theta} L(\pi(s_i; \theta))$;
  $d\theta_v \leftarrow d\theta_v (R - V(s_i; \theta_v))^2/2$;
end
Update θ with $d\theta$ and θ_v

Convergence of CI-CNN:

The entire learning process is documented in Algorithm 1.

Since the previous actor-critic network is hard to be optimized, we simplify the training process to guarantee the convergence. The key network training steps are summarized below:

(1) We simplify the training process to avoid the brittle convergence properties problem inherited from MTRL, we use the

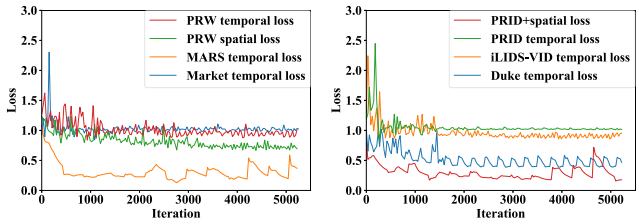


Fig. 7. Losses of the context-policy network.

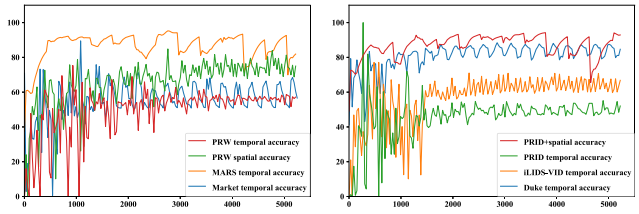


Fig. 8. Accuracy of the context-policy network.

network structure of CIN and copy the pre-trained parameters from the identification CIN to directly evaluate the context from the action space, of which, the 'context-critic' network is only upgraded on its last two layers. Therefore, the main parameters of the 'context-critic' are stable. Consequently, the loss of the 'context-critic' is consistent with the evaluation task for the actor network. This strategy reduces the complexity of the whole pipeline, and makes the context directly converge towards the ReID task.

(2) In terms of 'context-actor' network for the temporal and spatial context exploration tasks, we first alternatively train one of the two tasks separately with the other fixed, and then jointly train the two networks. In practice, we observe that, the temporal context task is much more easily converged than the spatial context. Therefore, we firstly train the temporal task, and then two networks are updated by the spatial context. Finally, the two tasks are jointly trained according to the evaluation from the 'context-critic' network.

(3) The 'context-actor' network in 'MTRL' model has a stable performance as shown in Fig. 7 and Fig. 8 during the training process. It is trained with our pre-trained model parameters from the CIN network. Moreover, we update the 'context-critic' network when needed. If the probability score given by 'context-critic' is slightly changed over different actions, it does not fit well with the context in the current environment, and we update the network based on Eq.4. This training process benefits the convergence robustness.

(4) The training of reinforcement learning network is postponed, such that it does not rely on the well-labeled datasets for context searching. This advantage alleviates the problem of dynamic context searching. Hence, the actor-critic network is superior over the directly back propagation optimization process.

IV. Experiments and Evaluations

A. Datasets and Experimental Settings

Our experiments are conducted on three publicly available video datasets, including PRW (Person Re-identification in the Wild) dataset [42], Market (Market1501) dataset [43], PRID (PRID2011) dataset [44], Duke (DukeMTMCT) dataset [45], MARS (MARS) dataset [46], iLIDS-VID dataset [29]. For person Re-ID, following the widely-used evaluation protocol from [29], we randomly split the dataset half-half for training/testing, and average the results of 10 rounds to make the evaluation stable. The splitting of probe/gallery in other three datasets all follow the provided settings in the original paper for fair comparison with others. The similarity is computed based on cosine metric. All of the six datasets preserve the original train/test split.

1) PRID: the PRID 2011 dataset [44] is collected in several outdoor scenes with relatively simple backgrounds. Videos in this dataset were captured by two static surveillance cameras. This dataset has an overlap of 200 pedestrians appeared in both views. Each person sequence has a variable length from 5 to 675 frames, with an average number of 100. To guarantee the effective length of image sequences, most previous works [8], [29] selected 178 identities with the sequence number more than 27 frames. On the contrast, our CI-CNN is adaptive to the flexible length of the sequences. The PRID dataset is split by the two different cameras, 'cam a' for training, 'cam b' for testing.

2) PRW: the PRW dataset [42] is captured with 6 cameras. It has 11,816 video frames, 932 identities belonging to 34,304 bounding boxes. To be noted that the identities in PRW do not correspond to those in Market dataset.

3) Market: the Market dataset is collected from six cameras. It contains 12,936 training images and 19,732 testing images. The number of identities is 751 and 750 in the training and testing sets respectively. There is an average of 17.2 images for each training identity. All the pedestrians are detected by the deformable part model (DPM) [47]. Both single and multiple query settings are used.

4) MARS: the MARS dataset [46] is a large-scale video dataset for person re-ID. It is captured from six near synchronized cameras in the Tsinghua campus. MARS contains 1,261 different identities, forming a total number of 20,478 tracklets. Among all the tracklets, there exist 3,278 distracted tracklets due to the false detection and association, making the dataset very challenging.

5) Duke: the Duke dataset [45] is collected from eight cameras. It contains 1,404 identities, of which, 702 identities for training and the remaining 702 identities for testing. The total training images are 16,522. In the testing set, for each identity, one query image per camera is picked up, and the remaining images are put in the gallery. There are 2,228 query images and 17,661 gallery images for the 702 testing identities.

6) iLIDS-VID: the ILIDS-VID dataset [29] is captured at an airport arrival hall. The dataset consists of 300 distinct individuals observed in two non-overlapping views, forming 600 pedestrian sequences. Each person has 23 to 192 images, and the average image number for each person is 73. It is very challenging because of the large clothing similarities among pedestrians, lighting and viewpoint variations across views, cluttered backgrounds and random occlusions.

Some examples of each datasets are shown in Fig. 9. We show the same identities from two different cameras in

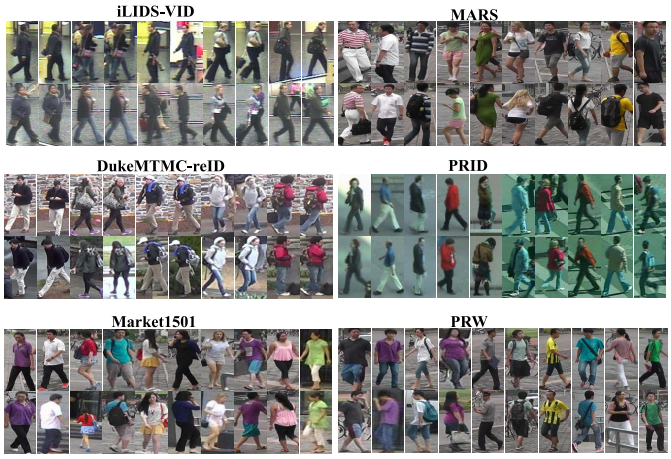


Fig. 9. Illustration of images in six benchmark datasets.

two rows. It can be seen that the pose, background and occlusions make it hard to recognize different pedestrians.

B. Implementation Details

We use the action set $\mathcal{A}$, which contains 9 actions to determine the spatial and temporal contexts. We initialize the context-policy network with a pre-trained Densenet-121 model [15], and finely-tune the fully-connected layers in two branches. One outputs spatial operations, the other outputs spatial operations. The number of the output units of the last layer is 3. The context-interactive network $e(s_t;\theta_e)$ and context-critic network $\pi(s_t;\theta)$ share the same CIN structure, except that the parameters of $e(s_t;\theta_e)$ updates 20 epoches earlier than $\pi(s_t;\theta)$. More concretely, our MTRL is implemented using Pytorch library on 4 Tesla P100 GPUs and one GTX1080 for testing the speed. The discount factor γ for the future rewards is set to 0.2.

All the networks are trained using stochastic gradient descent (SGD). On PRW, Market, Duke, iLIDS-VID, MARS, and PRID datasets, we use batch size of 32 for 60 epochs respectively. Following [15], the initial learning rate is set to be 0.1, and is divided by 10 at 50% and 75% of the total training epochs, we use a weight decay of 10^{-4} and a Nesterov momentum of 0.9 without dampening. We adopt the weight initialization introduced by [48]. For the three datasets without data augmentation, we add a dropout layer [49] after each convolutional layer (except the first one), and set the dropout rate to be 0.5. Specifically, to fairly compare with other methods, we employed the provided bounding boxes of the pedestrian in the original datasets to initialize the boxes.

C. Evaluation Settings

We use the generic evaluation metrics as previous works do [30]: top 1, 5, 10 and mAP (mean Average Precision). For each query, the average precision (AP) is computed according to its precision-recall curve. The mAP is calculated as the mean value of the average precisions of all queries. All the comparisons are based on single query. In particular, for iLIDS-VID dataset, we randomly split all the sequence pairs into two equal-size sets, with one for training and the other for testing. Then, we further select sequences from the first camera in the testing set to form the probe set, and those from the other camera are used as the gallery set. For the experiments on the MARS, Market, Duke, and PRW dataset, we use the reported single-query matching results. Two standard evaluation metrics are used: Cumulative Matching Characteristics (CMC) curve and Mean Average Precision (mAP). The CMC curve provides a ranking for each image in the gallery with respect to the probe. It is used for the 6 datasets.

D. Ablation Studies

We analyze the contribution of the main modules in our CI-CNN framework via an ablation experiment, including spatial context (SC), temporal context (TC), low rank analysis I_l in Eq. 6 (LR), and context-interactive network (CIN). We use the schemes of Densnet121, SC+ Densnet121, LR+GCN and TC+CIN on all of the three datasets. The Resnet50 and Densenet121 [15] are generally used as high performance baselines.

We observe that the aforementioned five modules will have fluctuant effects on different datasets, which is demonstrated in Table I. According to the results, the most contributive module is **TC+CIN** which has an improvement ranges from 25.66% (PRW) to 2.58% (Market) (top1 accuracy) by a large margin. The second most contributive module is LR+CIN, and the least contributive module is SC+Densenet121. It is rational due to the temporal context contains more visual clues than the spatial one. Besides, different datasets are improved by a slightly different margin. In fact, the PRW dataset has higher resolution than the Market dataset, which benefits the low-rank decomposition. Based on the results, we can draw conclusions: All the spatial context, temporal context, and the CIN can make a large margin compared with baselines. Of which, the temporal context contributes most, spatial context contributes least. Furthermore, the contributions of spatial context are dependent on the complexity of the scenarios.

There are two main advantages of the MTRL in this process. (1) MTRL can help search the temporal clues, which is further processed by low-rank decomposition. (2) MTRL can provide the initialized boxes highly related to the pedestrians' regions for CI-CNN to further extract the pedestrian's features from attention maps. More details are described in supplementary material.

E. Spatial and Temporal Parameter Analysis

To simplify the training process of CI-CNN, we design spatial and temporal context search actions in $\mathcal{A}$ with one scale each time via the grid search experiments to find the proper settings. We observe that the datasets can be classified into two classes: Temporal and spatial context based datasets, including PRID, PRW and Temporal context based datasets, including iLIDS-VID, Duke, Market and Mars. Therefore, rather than performing exhaustive experiments on the 6 datasets, we test on three datasets: PRW, PRID, and Market. The best setting is: zooming-out the image with 1.2 times and zooming-in the image with 0.8 times. When the current image requires more frames, 20 consecutive frames

TABLE I

PERFORMANCE COMPARISONS AMONG BASELINES AND OUR CI-CNN WITH DIFFERENT CONFIGURATIONS OVER THE PRW, MARKET, AND PRID2011 DATASETS. CIN DENOTES THE CONTEXT-INTERACTIVE NETWORK. THE ITALICS INDICATE THE SECOND-BEST PERFORMANCE

Dataset	PRW				Market				PRID			
Method	top1	top5	top10	mAP	top1	top5	top10	mAP	top1	top5	top10	mAP
ResNet50 [48]	38.30	50.51	55.97	16.59	88.30	91.92	95.25	84.26	66.32	71.27	75.89	49.21
Densenet121 [15]	39.28	50.81	56.21	19.23	90.12	92.12	95.82	84.21	75.98	78.39	82.36	62.86
LR+CIN	57.27	65.05	76.71	35.88	91.95	94.32	96.81	82.37	82.96	84.08	92.97	66.12
SC+Densenet121	50.70	51.79	66.16	22.81	91.24	92.98	97.12	86.12	79.61	81.95	86.74	64.83
TC+CIN	*64.94*	*70.46*	*74.64*	*43.96*	*92.70*	*93.21*	*97.21*	*86.21*	*84.23*	*88.52*	*91.23*	*72.45*
CI-CNN (ours)	**68.74**	**78.95**	**83.13**	**48.50**	**94.26**	**95.87**	**98.16**	**89.54**	**87.49**	**88.48**	**92.62**	**75.25**

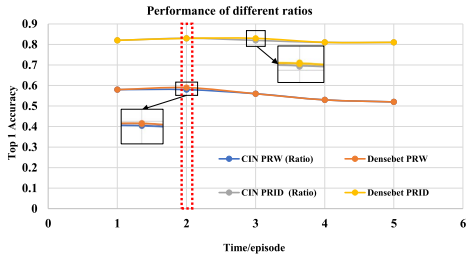


Fig. 10. The effects of ratio changes on top-1 accuracy over two datasets. It shows the close performance when the ratio is set from 0.1 to 0.5 times by step 0.1, per episode. The black boxes are zoomed in to see the details.

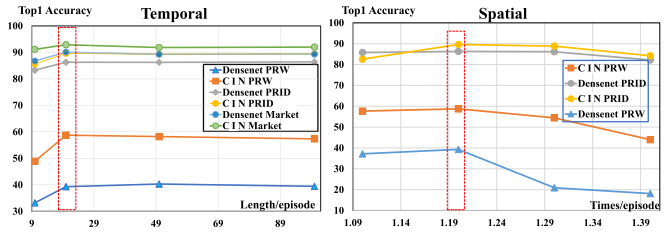


Fig. 11. The effects of temporal and spatial clues on top 1 accuracy over three datasets. It shows the best performance when the temporal frame number is set to 20 and the spatial setting is '1.2' times per episode.

will be provided, while the shrinking operation will reduce the neighboring frames by 10 frames. Thus, this setting is set as default in our method. To be noted that, the Market does not provide the original frames contains the pedestrians, therefore, we can only test the temporal context. The temporal curve in Fig. 11 shows that our CI-CNN is robust for the length of the temporal sequences, except that too few frames cannot provide temporal clues for person discrimination. Besides, the temporal curve also shows stable performance over different datasets. The spatial curve in Fig. 11 shows that the performance is sensitive to the spatial context which provides the person related regions across different scenarios. Zooming out the spatial region in a large margin (larger than 1.2 times) makes the performance decrease, however, it is also dependent on the dataset clutter extent of the background. Therefore, the adaptive spatial context search is necessary because the background is complex, which can not be pre-defined for person Re-ID.

Besides, we determine the ratios with the detection network [50] (which is the same as the detection network of PRW datasets), which tends to initialize the ratio with the semantic features of the persons.

To further explore the better search strategies of the ratios, we use progressive settings for the ratios. To be specific, the ratios are set from 0.1 to 0.5 by stepsize 0.1. The performance is shown in Fig. 10. The best performance is 0.2 times of the detected boxes, which sometimes causes decreasing in top-1 accuracy on PRID and PRW datasets.

In summary, the ratio search cannot improve the performance significantly, since the pedestrian ratio is close to the detection results, and the related context clues are in the nearby regions.

F. Comparisons With State-of-the-Art Methods

The following state-of-the-art person Re-ID methods are employed for comparison: video aggregation based methods

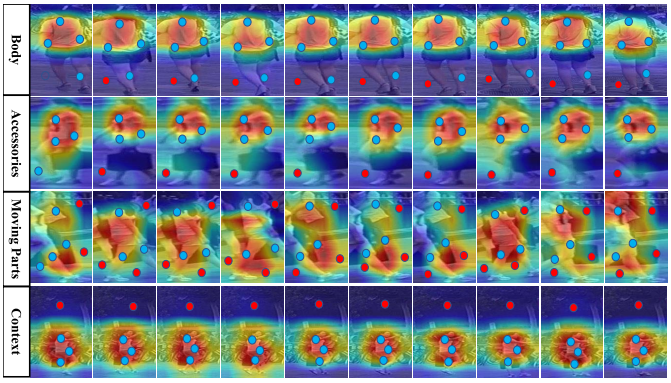


Fig. 12. Illustration of images in MARS dataset, the context and pedestrian attention map is extracted from the last convolution layer of CIN. Red dot is the context region, blue dot is the pedestrian region key point.

(e.g., [26], [30]), background removing methods (e.g., [2], [17], [51]). Please note that, video based methods use multi-shots, whose performances are relatively higher than those of the single-shot methods. We conduct comparisons based on the reported results or the results generated by the default-configuration networks on the same datasets. The results are shown in Table IV.

Performance: *PRW Dataset:* in this experiment, we compare with 7 state-of-the-art methods on PRW dataset. Spatial clues based methods do not use the temporal information, such as ResNet50 [48] with the strong discriminative feature. OIM [52], IAN [28], NPSM [27] and three temporal based methods utilize the temporal information, including DPM-Alex [42], ACF-Alex [42], LDCF+IDE [42]. It can be concluded that, our CI-CNN outperforms most of the state-of-the-art methods. The improvement is 6.89% (top1) on PRW dataset compared with the second-best performance achieved by IAN [28]. The PRW cannot performs as well as other larger datasets due to it has a smaller number (483 indentities) of training datasets.

TABLE II

COMPARISONS WITH STATE-OF-THE-ART METHODS OVER PRW DATASET

Method Category	PRW	top1	top5	top10	mAP
Spatial Methods	ResNet50 [48]	38.30	50.51	55.97	16.59
	OIM [52]	49.9	–	–	21.3
	IAN [28]	61.85	–	–	23.00
	NPSM [27]	53.1	–	–	24.2
Temporal Methods	DPM-Alex [42]	48.3	–	–	20.5
	ACF-Alex [42]	45.2	–	–	17.8
	LDCF+IDE [42]	45.5	–	–	18.3
Spatial+Temporal	**CI-CNN (Ours)**	**68.74**	78.95	83.13	48.50

TABLE III

COMPARISONS WITH STATE-OF-THE-ART METHODS OVER PRID DATASET. (BLUE ONES MEAN THAT IT CANNOT BE COMPARED DIRECTLY, BECAUSE THEY INVOLVE ADDITIONAL PRE-TRAINED DATASETS)

PRID	top1	top5	top10	mAP
ResNet50 [48]	66.32	71.27	75.89	49.21
CNN+RNN [53]	70	90	95	–
T-CN [10]	81.1	95.0	97.4	–
SeeForeset [6]	79.4	94.4	–	–
SPW [11]	83.5	96.3	–	–
QAN [26]	90.3	98.2	99.32	–
RL [30]	85.2	97.1	98.9	–
end AMOC+EpicFlow [54]	83.7	98.3	99.4	–
CI-CNN (Ours)	**87.49**	**98.48**	**99.62**	**75.25**

TABLE IV

COMPARISONS WITH STATE-OF-THE-ART METHODS OVER MARKET DATASET

Method Category	Method	top1	top5	top10	mAP
None-Attention	Densenet121 [15]	85.3	93.3	95.8	65.0
	Basel (D,Tri) [16]	86.73	–	–	67.78
	PSE+ECN [20]	90.3	–	–	84
	Part-Aligned [55]	93.4	96.4	97.4	89.9
	BPM [56]	93.8	97.5	98.5	81.6
Attention Methods	Background [5]	81.2	94.6	97	—
	DuATM [17]	91.42	97.09	–	76.62
	AACN [2]	88.69	–	–	82.96
	ACS [51]	88.66	–	–	83.3
	CI-CNN (Ours)	**94.23**	**97.81**	**99.27**	**83.93**

TABLE V

COMPARISONS WITH STATE-OF-THE-ART METHODS OVER iLIDS-VID DATASET

Method	top1	top5	top10	mAP
T-CN [10]	60.6	83.8	91.2	–
SeeForeset [6]	55.2	86.5	–	–
end AMOC+EpicFlow [54]	68.7	94.3	98.3	–
QAN [26]	68	86.8	95.4	–
RL [30]	60.2	84.7	91.7	–
Densenet121 [15]	54.27	62.19	68.13	37.89
CI-CNN (Ours)	**71.23**	**80.25**	**85.81**	**40.15**

PRID Dataset: in this experiment, we compare our CI-CNN with ResNet50 [48], T-CN [10], SeeForeset [6], SPW [11], QAN [26], RL [30]. The performance is documented in Table III.

The performance gain of our CI-CNN on PRID dataset ranks last, because the PRID dataset has simple background, no occlusion, and small pose variance. However, compared with other datasets, the performance baseline on PRW dataset is slightly lower, because the number of persons in the training dataset is only 483, each person has less than 20 frames on average while involving 6 camera views, and the high accuracy in the training scenarios may easily cause overfitting. In summary, our CI-CNN has a dramatic improvement across different datasets with various scenarios.

Market Dataset: we evaluate on the large dataset market, wherein the training set contains 12936 bounding boxes of 750 identities. The testing set contains 19732 bounding boxes of 751 identities. Compared to the above three small size datasets iLIDS-VID, PRID and PRW, Market-1501 has more complex intra-class and inter-class variations. The market dataset is compared with none-attention methods: Densenet121 [15], Basel (D,Tri) [16], PSE+ECN [20], Part-Aligned [55], BPM [56], and attention methods: Background [5], DuATM [17], AACN [2], ACS [51]. The largest improvement on Market dataset is 2.81 % (top1) compared with the second-best performance achieved by DuATM [17].

On the other hand, to compare with the state-of-the-art methods, we conduct experiments on the widely-used iLIDS-VID, MARS and Duke video datasets. Since no spatial context is provided in the dataset, we extract the temporal contexts in image sequences ordered by the cameras. The length is determined by the MTRL. Note that in the cross scenarios datasets, we exclude 253 IDs in Market dataset which are overlapped with the PRW datasets.

iLIDS-VID Dataset: the iLIDS-VID dataset is compared with the state-of-the-art works, including T-CN [10], AMOC+EpicFlow [54], SeeForeset [6], PAM-LOMO+KISSME [57], QAN [26], DRSA [58], RL [30] methods. These compared methods are mostly based on the temporal sequences, which have taken advantages of sufficient temporal information to achieve better performance w.r.t single image based methods.

According to Table VII, it can be concluded that our CI-CNN achieves the comparable performance with state-of-the-art methods. The iLIS-VID dataset has stable background and spatial context, the improvement mainly benefits from the temporal context clues. wherein, the total improvement is 7.02% in top 1 accuracy.

MARS Dataset: the MARS dataset is compared with the state-of-the-art works, including MSCAN [59], Part-Aligned [55], BPM [56], DuATM [17], ETAP-Net [58]. The results are shown in Table VI. It shows that, our CI-CNN outperforms other state-of-the-art method by a large margin. Specially, the improvement of mAP is 2.7% compared with the second best method BPM [56]. The MARS dataset have the largest number of candidate persons compared with the other 5 datasets. The background context is the most complex due to the various places, multiple persons, vehicles and occlusions. We visualize the context attention map from the last convolution layer of our CI-CNN. It is obvious that, our CI-CNN dynamically determine the context region for specific pedestrian image, as shown in Fig. 12. For the easy samples with clear pedestrian and simple background, the context will focus on the body part, which is the most discriminative and largest region for persons. On the contrast, when the target pedestrian is hard to be recognized, it will focus on the whole torso of the person with a higher probability, while the surroundings with a lower probability.

TABLE VI

COMPARISONS WITH STATE-OF-THE-ART METHODS OVER MARS DATASET

Method	top1	top5	top10	mAP
MSCAN [59]	83.03	93.69	–	66.43
Part-Aligned [55]	85.1	94.2	97.4	83.9
BPM [56]	86.3	94.7	98.2	76.1
DuATM [17]	78.74	90.86	–	62.26
ETAP-Net [58]	80.75	92.07	–	67.39
SeeForeset [6]	70.6	90.0	–	50.7
Densenet121 [15]	74.51	76.53	77.37	63.48
CI-CNN (Ours)	**87.3**	**93.7**	**98.4**	**78.8**

TABLE VII

COMPARISONS WITH STATE-OF-THE-ART METHODS OVER DUKE DATASET

Method Category	Method	top1	top5	top10	mAP
None-Attention	Densenet121 [15]	73.6	85.3	88.4	52.4
	Basel.+LSRO [60]	67.68	–	–	47.13
	Basel.+OIM [61]	68.1	–	–	–
	ACRN [62]	72.58	84.79	88.87	51.96
Attention Methods	SVDNet [63]	76.7	86.4	89.9	56.8
	Chen et. al. [64]	79.2	–	–	60.6
	DuATM [17]	81.62	84.79	–	64.58
	AACN [2]	76.84	–	–	59.25
	ACS [51]	84.11	–	–	78.19
	Basel (D,Tri) [16]	77.03	–	–	55.34
	PSE+ECN [20]	*85.2*	–	–	*79.8*
	FD-GAN [65]	79.90	88.6	90.8	64.5
	CI-CNN (Ours)	**87.6**	**91.2**	**94.5**	**81.3**

Duke Dataset: the Duke dataset is compared with the state-of-the-art works. There are two main streams: none-attention based methods including Densenet121 [15], Basel.+LSRO [60], Basel.+OIM [61], ACRN [62], SVDNet [63], and attention based methods including Chen et. al. [64], DuATM [17], AACN [2], ACS [51], Basel (D,Tri) [16], PSE+ECN [20]. As shown in Table VII, the improvement compared with the second best method is 2.4%. The improvement comes from the temporal context.

Efficiency: Compared with the previous temporal clues dependent works, our network costs the least memory and time in the inference phase. Specially, to involve the sequences of 100 frames, the memory cost of the LSTM [66] network will be linearly increased 100 times w.r.t the single image cases. On the contrast, ours will go through an actor-critic framework, which contains a single network and a single CIN network. The memory cost is twice of the Densenet121. The memory cost of the CI-CNN network is also stable with the length increase of the image sequences, which is superior to conventional LSTM based methods (e.g., SeeForest [6]). The memory and computing complexity are documented in Table VIII. It can be concluded that our CI-CNN has the least cost compared with the temporal information based methods.

G. Cross-Scene Validation

We conduct three kinds of experiments on the 6 datasets to demonstrate our CI-CNN's unsupervised transfer learning ability. *1) Transferring ability across datasets.* We test on PRW, PRID, and Market datasts. All the models of CI-CNN's are trained only over the source dataset, and then are used to conduct unsupervised feature extraction over the target

TABLE VIII

EFFICIENCY COMPARISONS WITH STATE-OF-THE-ART METHODS

Methods	FlOPS (GMac)	Params (M)	FPS	Memory (MB)	Memory ratio	Training (hours)
Densnet121	2.88	7.0	10	2132	30	10
Resnet50	5.19	25.05	100	256	4.21	14
Camera	4.12	28.0	20	603	7.43	15
Resnet503d [67]	–	47.48	–	12227	75.12	20
resnet50rnn [53]	–	24.78	–	7577	47.66	20
SeeForeset [6]	–	–	20	665	8.2	3
CI-CNN (Ours)	2.92	7.0	5	2264	30	12

TABLE IX

COMPARISONS OF UNSUPERVISED TRANSFER LEARNING ABILITY. ALL THE MODELS ARE TRAINED ONLY ON SOURCE DATASET

Method	Source	Target	Performance			
			top1	top5	top10	mAP
Resnet50	PRW	Market	52.70	65.56	70.96	35.93
	PRID	Market	41.25	52.31	55.83	24.54
	Market	PRW	89.22	94.63	96.86	85.12
	PRID	PRW	21.35	24.16	28.92	14.49
	PRW	PRID	24.15	31.98	34.17	17.81
	Market	PRID	32.18	37.51	42.22	19.38
Densenet121	PRW	Market	85.54	91.33	93.44	79.66
	PRID	Market	45.23	55.56	60.96	25.93
	Market	PRW	92.64	96.14	97.15	89.06
	PRID	PRW	26.84	31.76	35.13	14.19
	PRW	PRID	29.35	38.91	44.57	20.46
	Market	PRID	41.34	46.31	49.05	24.85
CI-CNN(ours)	**PRW**	**Market**	**85.94**	**92.18**	**95.65**	**81.03**
	PRID	Market	61.45	67.33	70.52	34.52
	Market	PRW	**93.24**	**95.31**	**97.16**	**84.39**
	PRID	PRW	42.47	53.21	56.31	25.27
	PRW	PRID	54.36	62.14	67.47	32.42
	Market	PRID	48.81	52.63	58.14	32.25
TJ-AIDJ [32]	Duke	Market	58.2	74.8	81.1	26.5
	Market	PRID	26.8	–	–	–
FD-GAN [65]	Market	PRW	67.0	82.6	87.3	65.9
	Market	PRID	21.4	34.2	37.2	15.8
Camera [68]	Market	PRW	82.3	89.5	90.8	76.4
	Market	PRID	21.4	34.2	37.2	15.8

dataset, shown in Table IX. In all testing cases, our CI-CNN outperforms the baseline networks by a large margin. For example, when PRW serves as target dataset and Market as source dataset, the performance is improved largest according to our transfer experiments. It may be attributed to that the source dataset market has larger dataset than PRW while having similar scenarios. Therefore, our CI-CNN can sufficiently extract common context on Market dataset and PRW dataset, which contains much more sequences across different scenarios. The results show our method gives rise to performance improvement by a large margin compared with the baselines and the existing state-of-the-art methods.

2) Comparisons with state-of-the-art methods. We also compare our method with the state-of-the-art unsupervised cross-dataset person Re-ID algorithm [32], FD-GAN [65], Camera [68]. All the methods addresse the similar problem with us under the same dataset configuration, and document the results in Table IX. In all the testing cases, our CI-CNN outperforms TJ-AIDJ by a large margin. For example, the TJ-AIDJ is pre-trained on the Duke datasets, which has more diverse backgrounds and samples than PRID. It is the reason to achieve high performance in terms of top5 and top10 indicators. Specially, for the cases with PRID as the target dataset, our CI-CNN performs extremely well (3.25%

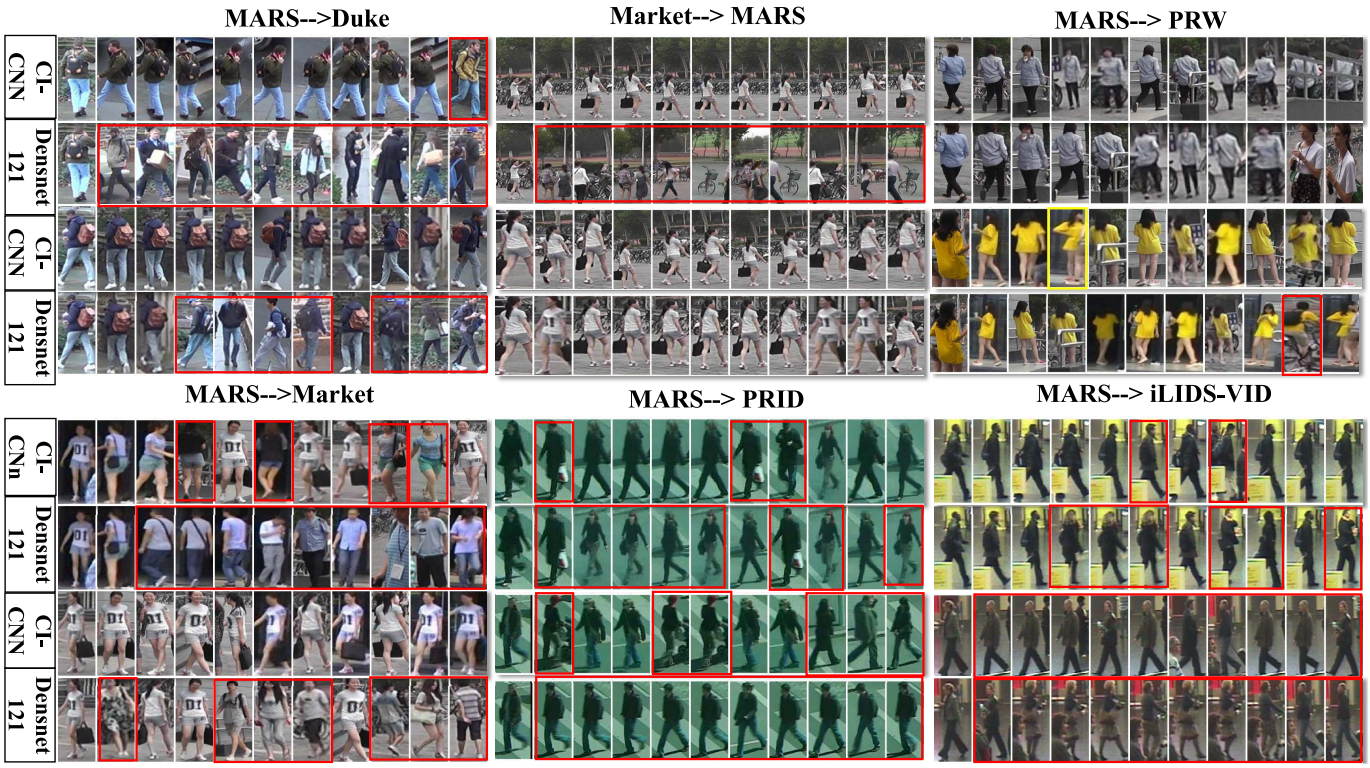


Fig. 13. Illustration of query and gallery images, the first is the query image, following ranked galleries based on top10 indicator. The the red boxes denote the falsely recognized cases (we do not exclude the same-camera dataset w.r.t the query image, we exclude the same-camera captured same persons in the CMC metric).

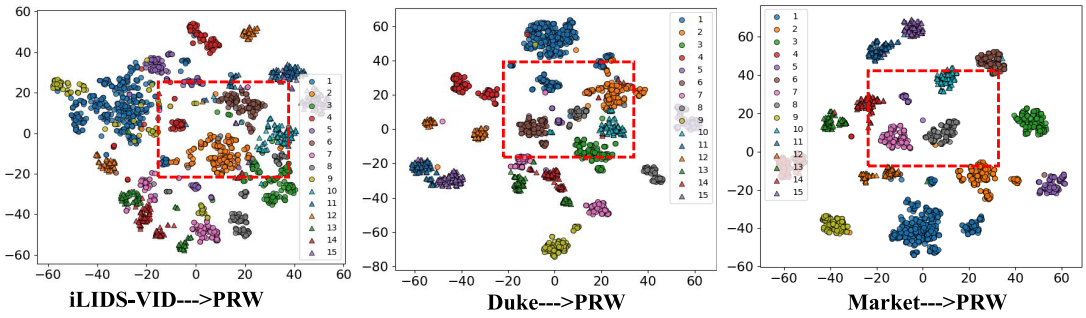


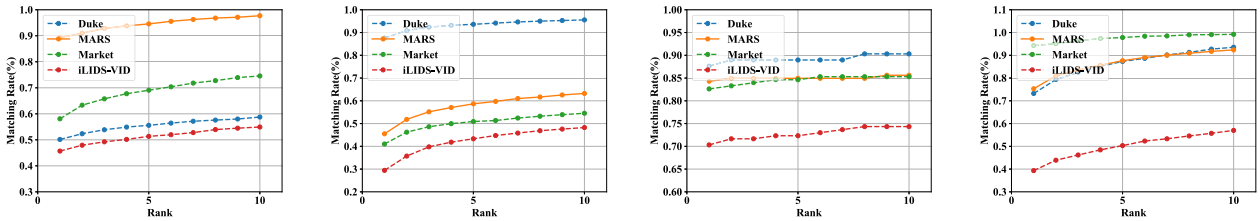
Fig. 14. Illustration of generalization ability of our CI-CNN on pre-trained datasets, including iLIDS-VID, Duke, and Market.

higher on top1 accuracy). Therefore, our CI-CNN has the superior ability in adaptively searching spatial and temporal contexts, and can significantly improve the Re-ID performance. Moreover, to study the improved cases of our CI-CNN, we list some examples on the query and top10 gallery images in Fig. 13. The results show that, our CI-CNN can benefit from the context region for the cases with complex background.

Due to the Duke and Market dataset have non-overlapped samples, most state-of-the-art domain-adaption methods are trained on Duke and tested on Market dataset. We compare with the newly published methods including, Wang *et al.* [32], Wei *et al.* [69], Deng *et al.* [70], Wei *et al.* [69], Zhong *et al.* [71]. The results are shown in Table X. Our CI-CNN yields best performance on Market dataset. 'Duke' means the source dataset is Duke, 'Duke labels' means Duke labels are used for domain adaptation.

3) Generalization on multiple scenarios: we further study the generalization ability of our CI-CNN on different datasets. We use one of Mars, ILIDS-VID, Duke and Market as the training dataset and test on the other three datasets. The iLIDS-VID dataset involves the most different scenes w.r.t the PRW dataset, while the Market dataset is the closest to the PRW datast in terms of scene diversity. As shown in Fig. 14, with the similar context increasing in the training dataset, the performance will also increase on the testing dataset. The full cross scenarios comparison results are shown in Fig. 15.

In conclusion, the context information has two advantages in transferring the common features across scenarios. (1) Searching the spatial-temporal clues, of which, the searching strategy is stable across scenarios for the target pedestrians; (2) Transferring the attention maps to indicate the context relations related to the target persons across scenarios. This is further guaranteed by the attention maps trained in the network.



(a) Mars Dataset (b) Duke Dataset (c) iLIDS-VID Dataset (d) Market Dataset

Fig. 15. CMC curves of the 4 datasets without the spatial context.

TABLE X

COMPARISONS OF UNSUPERVISED TRANSFER LEARNING ABILITY. ALL THE MODELS ARE TRAINED ONLY ON DUKE DATASET, AND ARE THEN USED TO CONDUCT UNSUPERVISED FEATURE EXTRACTION ON THE MARKET DATASET

Method	top1	top5	top10	mAP	Source Domain
Wang el al. [32]	58.2	74.8	81.1	26.5	Duke
Wei et al. [69]	38.6	-	66.1	-	Duke
SPGAN [70]	51.5	70.1	76.8	22.8	Duke labels
Wei et al. [69]	33.5	-	61.9	-	Duke
Zhong et al. [71]	62.2	78.8	84	31.4	Duke
Camera [68]	43.3	62.6	69.7	18.0	Duke
Densenet121 [15]	43.3	62.6	69.7	18.0	Duke
CI-CNN (Ours)	**73.2**	**81.8**	**87.2**	**47.3**	Duke

This attention map is stable in pedestrians' images with the variances of pose, view, and clutter backgrounds. We provide visualized results in supplementary material.

V. CONCLUSION AND FUTURE WORK

In this paper, we have presented CI-CNN as a high-performance cross-scenario person Re-ID framework. In particular, our CI-CNN could transfer the context across different camera views by adaptively finding the spatial-spatial context and decomposing the context using low-rank analysis. In addition, a context-interactive network is proposed to enhance the person-related spatial feature extraction. Experiments show that our CI-CNN outperforms most of the state-of-the-art person Re-ID methods on the scene-independent datasets by a large margin. In the near future, we shall extend our newly-proposed CI-CNN to handle other real-world tasks in more complex environments such as raining day and night time. In addition, we will extend our CI-CNN to action recognition or pose estimation tasks by taking advantage of the motion features in critical video analysis applications.

REFERENCES

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

[2] J. Xu, R. Zhao, F. Zhu, H. Wang, and W. Ouyang, "Attention-aware compositional network for person re-identification," in *Proc. CVPR*, Jun. 2018, pp. 2119–2128.

[3] Z. D. Kai Li, "Discriminative semi-coupled projective dictionary learning for low-resolution person re-identification," in *Proc. AAAI*, Jan. 2018, pp. 2331–2338. [Online]. Available: http://par.nsf.gov/biblio/10065429

[4] 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.

[5] M. Tian *et al.*, "Eliminating background-bias for robust person re-identification," in *Proc. CVPR*, 2018, pp. 5794–5803.

[6] 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 *Proc. CVPR*, 2017, pp. 6776–6785.

[7] X. Chang, P.-Y. Huang, Y.-D. Shen, X. Liang, Y. Yang, and A. G. Hauptmann, "RCAA: Relational context-aware agents for person search," in *Proc. ECCV*, 2018, pp. 84–100.

[8] J. Dai, P. Zhang, D. Wang, H. Lu, and H. Wang, "Video person re-identification by temporal residual learning," *IEEE Trans. Image Process.*, vol. 28, no. 3, pp. 1366–1377, Mar. 2019.

[9] X. Zhu, X.-Y. Jing, X. You, X. Zhang, and T. Zhang, "Video-based person re-identification by simultaneously learning intra-video and inter-video distance metrics," *IEEE Trans. Image Process.*, vol. 27, no. 11, pp. 5683–5695, Nov. 2018.

[10] Y. Wu, J. Qiu, J. Takamatsu, and T. Ogasawara, "Temporal-enhanced convolutional network for person re-identification," in *Proc. AAAI*, 2018, pp. 7412–7419.

[11] W. Huang, C. Liang, Y. Yu, Z. Wang, W. Ruan, and R. Hu, "Video-based person re-identification via self paced weighting," in *Proc. AAAI*, 2018, pp. 2273–2280.

[12] L. Wu, Y. Wang, J. Gao, and X. Li, "Where-and-when to look: Deep Siamese attention networks for video-based person re-identification," *IEEE Trans. Multimedia*, vol. 21, no. 6, pp. 1412–1424, Jun. 2019.

[13] J. Tang, X. Shu, R. Yan, and L. Zhang, "Coherence constrained graph LSTM for group activity recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published.

[14] X. Shu, J. Tang, G.-J. Qi, W. Liu, and J. Yang, "Hierarchical long short-term concurrent memory for human interaction recognition," 2018, *arXiv:1811.00270*. [Online]. Available: https://arxiv.org/abs/1811.00270

[15] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. CVPR*, Jul. 2017, pp. 2261–2269.

[16] J. Liu, B. Ni, Y. Yan, P. Zhou, S. Cheng, and J. Hu, "Pose transferrable person re-identification," in *Proc. CVPR*, 2018, pp. 4099–4108.

[17] J. Si *et al.*, "Dual attention matching network for context-aware feature sequence based person re-identification," in *Proc. CVPR*, Jun. 2018, pp. 5363–5372.

[18] B. Nguyen and B. De Baets, "Kernel distance metric learning using pairwise constraints for person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 2, pp. 589–600, Feb. 2019.

[19] G. Liang, X. Lan, K. Zheng, S. Wang, and N. Zheng, "Cross-view person identification by matching human poses estimated with confidence on each body joint," in *Proc. AAAI*, 2018, pp. 7089–7097.

[20] M. S. Sarfraz, A. Schumann, A. Eberle, and R. Stiefelhagen, "A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking," in *Proc. CVPR*, Jun. 2018, pp. 420–429.

[21] Y.-J. Cho and K.-J. Yoon, "PaMM: Pose-aware multi-shot matching for improving person re-identification," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3739–3752, Aug. 2018.

[22] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "CamStyle: A novel data augmentation method for person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 3, pp. 1176–1190, Mar. 2019.

[23] Z. Feng, J. Lai, and X. Xie, "Learning view-specific deep networks for person re-identification," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3472–3483, Jul. 2018.

[24] A. Borgia, Y. Hua, E. Kodirov, and N. M. Robertson, "Cross-view discriminative feature learning for person re-identification," *IEEE Trans. Image Process.*, vol. 27, no. 11, pp. 5338–5349, Nov. 2018.

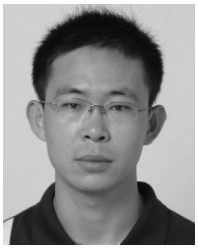
[25] M. M. Kalayeh, E. Basaran, M. Gökmen, M. E. Kamasak, and M. Shah, "Human semantic parsing for person re-identification," in *Proc. CVPR*, 2018, pp. 1062–1071.

[26] Y. Liu, J. Yan, and W. Ouyang, "Quality aware network for set to set recognition," in *Proc. CVPR*, 2017, pp. 4694–4703.

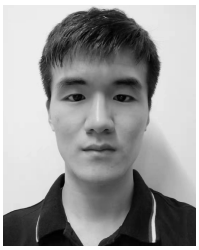
[27] H. Liu *et al.*, "Neural person search machines," in *Proc. ICCV*, 2017, pp. 493–501.

[28] J. Xiao, Y. Xie, T. Tillo, K. Huang, Y. Wei, and J. Feng, "IAN: The individual aggregation network for person search," 2017, *arXiv:1705.05552*. [Online]. Available: https://arxiv.org/abs/1705.05552
[29] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person re-identification by video ranking," in *Proc. ECCV*, 2014, pp. 688–703.
[30] J. Zhang, N. Wang, and L. Zhang, "Multi-shot pedestrian re-identification via sequential decision making," in *Proc. CVPR*, Jun. 2018, pp. 6781–6789.
[31] J. Lv, W. Chen, Q. Li, and C. Yang, "Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns," in *Proc. CVPR*, 2018, pp. 7948–7956.
[32] J. Wang, X. Zhu, S. Gong, and W. Li, "Transferable joint attribute-identity deep learning for unsupervised person re-identification," in *Proc. CVPR*, Jun. 2018, pp. 2275–2284.
[33] J. Han, L. Yang, D. Zhang, X. Chang, and X. Liang, "Reinforcement cutting-agent learning for video object segmentation," in *Proc. CVPR*, 2018, pp. 9080–9089.
[34] X. Wang, W. Chen, J. Wu, Y.-F. Wang, and W. Y. Wang, "Video captioning via hierarchical reinforcement learning," in *Proc. CVPR*, 2018, pp. 4213–4222.
[35] D. Li, H. Wu, J. Zhang, and K. Huang, "A2-RL: Aesthetics aware reinforcement learning for image cropping," in *Proc. CVPR*, 2018, pp. 8193–8201.
[36] S. Lan, R. Panda, Q. Zhu, and A. K. Roy-Chowdhury, "FFNet: Video fast-forwarding via reinforcement learning," in *Proc. CVPR*, Jun. 2018, pp. 6771–6780.
[37] B. Zhong, B. Bai, J. Li, Y. Zhang, and Y. Fu, "Hierarchical tracking by reinforcement learning based searching and coarse-to-fine verifying," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2331–2341, May 2019.
[38] Z. Yang, K. E. Merrick, H. A. Abbass, and L. Jin, "Multi-task deep reinforcement learning for continuous action control," in *Proc. IJCAI*, 2017, pp. 3301–3307.
[39] V. Mnih *et al.*, "Asynchronous methods for deep reinforcement learning," in *Proc. ICML*, 2016, pp. 1928–1937.
[40] V. R. Konda and J. N. Tsitsiklis, "Actor-critic algorithms," in *Proc. NIPS*, 2000, pp. 1008–1014.
[41] T. Bouwmans and E. H. Zahzah, "Robust PCA via principal component pursuit: A review for a comparative evaluation in video surveillance," *Comput. Vis. Image Understand.*, vol. 122, pp. 22–34, May 2014.
[42] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian, "Person re-identification in the wild," 2016, *arXiv:1604.02531*. [Online]. Available: https://arxiv.org/abs/1604.02531
[43] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. CVPR*, 2015, pp. 1116–1124.
[44] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," in *Proc. SCIA*, 2011, pp. 91–102.
[45] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Proc. ECCVW*, 2016.
[46] L. Zheng *et al.*, "MARS: A video benchmark for large-scale person re-identification," in *Proc. ECCV*, 2016, pp. 868–884.
[47] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.
[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.
[49] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.
[50] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. NIPS*, 2015, pp. 91–99.
[51] H. Huang, D. Li, Z. Zhang, X. Chen, and K. Huang, "Adversarially occluded samples for person re-identification," in *Proc. CVPR*, Jun. 2018, pp. 5098–5107.
[52] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "Joint detection and identification feature learning for person search," in *Proc. CVPR*, 2017, pp. 3376–3385.
[53] N. McLaughlin, J. M. del Rincon, and P. Miller, "Recurrent convolutional network for video-based person re-identification," in *Proc. CVPR*, 2016, pp. 1325–1334.
[54] H. Liu *et al.*, "Video-based person re-identification with accumulative motion context," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 2788–2802, Oct. 2018.
[55] Y. Suh, J. Wang, S. Tang, T. Mei, and K. Mu Lee, "Part-aligned bilinear representations for person re-identification," in *Proc. ECCV*, 2018, pp. 402–419.
[56] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proc. ECCV*, 2018, pp. 480–496.
[57] F. M. Khan and F. Brèmond, "Multi-shot person re-identification using part appearance mixture," in *Proc. WACV*, 2017, pp. 605–614.
[58] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Ouyang, and Y. Yang, "Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning," in *Proc. CVPR*, 2018, pp. 5177–5186.
[59] 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. CVPR*, 2017, pp. 384–393.
[60] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by GAN improve the person re-identification baseline *in vitro*," in *Proc. ICCV*, Oct. 2017, pp. 3754–3762.
[61] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "Joint detection and identification feature learning for person search," in *Proc. CVPR*, 2017, pp. 3415–3424.
[62] A. Schumann and R. Stiefelhagen, "Person re-identification by deep learning attribute-complementary information," in *Proc. CVPRW*, 2017, pp. 1435–1443.
[63] Y. Sun, L. Zheng, W. Deng, and S. Wang, "SVDNet for pedestrian retrieval," in *Proc. ICCV*, 2017, pp. 3800–3808.
[64] 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.
[65] Y. Ge *et al.*, "FD-GAN: Pose-guided feature distilling GAN for robust person re-identification," in *Proc. NIPS*, 2018, pp. 1229–1240.
[66] J. Gao and R. Nevatia, "Revisiting temporal modeling for video-based person ReID," 2018, *arXiv:1805.02104*. [Online]. Available: https://arxiv.org/abs/1805.02104
[67] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet?" in *Proc. CVPR*, 2018, pp. 6546–6555.
[68] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camera style adaptation for person re-identification," in *Proc. CVPR*, 2018, pp. 5157–5166.
[69] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer GAN to bridge domain gap for person re-identification," in *Proc. CVPR*, 2018, pp. 79–88.
[70] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao, "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification," in *Proc. CVPR*, 2018, pp. 994–1003.
[71] Z. Zhong, L. Zheng, S. Li, and Y. Yang, "Generalizing a person retrieval model hetero- and homogeneously," in *Proc. ECCV*, 2018, pp. 172–188.

Wenfeng Song received the M.S. degree in software engineering from Beihang University, Beijing, China, in 2015, where she is currently pursuing the Ph.D. degree in technology of computer application. Her research interests include pattern recognition, computer vision, and machine learning.



Shuai Li received the Ph.D. degree in computer science from Beihang University. He is currently an Associate Professor with the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University. His research interests include computer graphics, pattern recognition, computer vision, and medical image processing.



Tao Chang received the B.E. degree in computer science from Beihang University, Beijing, China, in 2018, where he is currently pursuing the M.S. degree in computer science and Technology. His research interests include computer vision and machine learning.

Aimin Hao received the B.S., M.S., and Ph.D. degrees in computer science from Beihang University. He is currently a Professor with the Computer Science School and the Associate Director of the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University. His research interests are on virtual reality, computer simulation, computer graphics, geometric modeling, image processing, and computer vision.

Qinping Zhao received the Ph.D. degree in computer science from Nanjing University, in 1989. He is currently a Professor with Beihang University (BUAA), a member of the China Academy of Engineering (CAE), the Chief Scientist of State Key Laboratory of Virtual Reality Technology and System, and the President of China Simulation Federation (CSF). His research interests are on virtual reality, computer simulation, and computer vision.

Hong Qin received the B.S. and M.S. degrees in computer science from Peking University, the Ph.D. degree in computer science from the University of Toronto. He is currently a Professor of computer science with the Department of Computer Science, Stony Brook University. His research interests include geometric and solid modeling, graphics, physics-based modeling and simulation, computer-aided geometric design, visualization, and scientific computing.