Temporal Complementarity-Guided Reinforcement Learning for Image-to-Video Person Re-Identification

Wei Wu†, Jiawei Liu†, Kecheng Zheng, Qibin Sun, Zheng-Jun Zha*
University of Science and Technology of China, China
{wuvy,zkcys001}@mail.ustc.edu.cn, {jwliu6,qibinsun,zhazj}@ustc.edu.cn

Abstract

Image-to-video person re-identification aims to retrieve the same pedestrian as the image-based query from a video-based gallery set. Existing methods treat it as a cross-modality retrieval task and learn the common latent embeddings from image and video modalities, which are both less effective and efficient due to large modality gap and redundant feature learning by utilizing all video frames. In this work, we first regard this task as point-to-set matching problem identical to human decision process, and propose a novel Temporal Complementarity-Guided Reinforcement Learning (TCRL) approach for image-to-video person re-identification. TCRL employs deep reinforcement learning to make sequential judgments on dynamically selecting suitable amount of frames from gallery videos, and accumulate adequate temporal complementary information among these frames by the guidance of the query image, towards balancing efficiency and accuracy. Specifically, TCRL formulates point-to-set matching procedure as Markov decision process, where a sequential judgement agent measures the uncertainty between the query image and all historical frames at each time step, and verifies that sufficient complementary clues are accumulated for judgment (same or different) or one more frames are requested to assist judgment. Moreover, TCRL maintains a sequential feature extraction module with complementary residual detectors to dynamically suppress redundant salient regions and thoroughly mine diverse complementary clues among these selected frames for enhancing frame-level representation. Extensive experiments demonstrate the superiority of our method.

1. Introduction

Person re-identification (Re-ID) is the task of searching the target samples from the gallery set which have the same identity with the given query. [5, 35]. It has been widely studied in the computer vision community during the past few years, due to its large potential in the intelligent surveillance, video analysis and human-robot interaction, *etc* [5, 42]. It is a quite challenging task, derived from the dramatic variations in camera viewpoint, body pose, illumination, as well as the influence of cluttered background and partial occlusion.

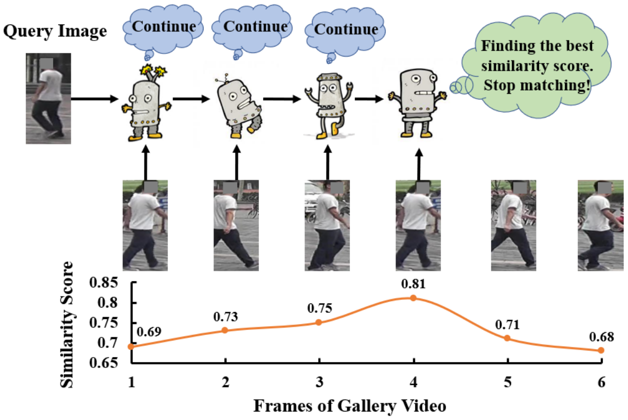


Figure 1. Image-to-video person Re-ID essentially belongs to point-to-set matching problem, identical to human decision process for pedestrian matching. Due to abundant redundant information and even noisy information within consecutive frames, the best similarity score between a query image and a gallery video is often not obtained when employing all frames of the video.

In general, person Re-ID can be mainly divided into two categories: image-based Re-ID [8, 19, 43] and video-based Re-ID [3, 36, 39]. The main difference between them is that for the former, the query and gallery samples are both images, while the query and gallery samples are both videos for the latter. In these two categories, the samples to be matched are homogeneous. Recently, image-based and video-based Re-ID have achieved impressive progress, benefiting from the development of deep learning technique. However, in many practical scenarios, person Re-ID requires to find the target pedestrian in numerous videos according to a query image. One situation is that, given an im-

† Equal contribution.
* Corresponding author.

age of a criminal, the person Re-ID system should retrieve the criminal across multiple non-overlap video sequences. Such brings out an emerging task, *i.e.*, image-to-video (I2V) person re-identification [7, 28, 33].

Contrary to image and video based Re-ID, I2V Re-ID is more challenging due to the information asymmetry between images and videos. Videos contain plenty of temporal information across time dimension, which results in feature distribution discrepancy and increasing the difficulty of measuring the similarity scores between image and video samples [25]. To tackle with this issue, existing I2V Re-ID methods dedicate to *1) project images and videos into a common embedding space by distance metric learning* [27, 30, 33, 47, 48] or *2) propagate the temporal knowledge learned from the video representation network to the image representation network via temporal knowledge distillation* [7, 25, 28]. However, the aforementioned methods are both less effective and efficient. They simply treat I2V Re-ID as a cross-modality retrieval task, and enforce image and video features to resemble each other even though the image is completely different from the video due to lacking temporal dimension. Moreover, as illustrated in Figure 1, video sequences often mingle vast redundant appearance clues and noisy information, these methods directly exploit all frames of videos without discovering discriminative complementary information among them and avoiding the interference of noisy information, leading to poor feature representation and inefficient model.

In this work, we first regard this task as point-to-set matching problem that is the same with human decision process, and propose a novel Temporal Complementarity-Guided Reinforcement Learning (TCRL) approach for image-to-video person re-identification. TCRL exploits deep reinforcement learning to accumulate adequate complementary information from suitable amount of frames by making sequential judgements on the query images and the gallery videos, towards balancing efficiency and accuracy. Concretely, TCRL formulates point-to-set matching procedure as Markov decision process, where a sequential judgement agent measures the uncertainty between the image feature and the frame-level feature containing the temporal complementary information of all historical frames at each time step, and learns the optimal policy to make the judgment that the model have collected enough evidence for identifying the same pedestrian and distinguishing different pedestrians, or requests one more video frames to assist recognizing. In addition, TCRL designs a sequential feature extraction module with complementary residual detectors to improve the capacity of the frame-level feature by absorbing the diverse complementary information among these selected video frames. The complementary residual detector learns the most salient features that have been activated in previous frames by a multi-head attention mechanism, which are then utilized as the salient convolutional kernel to estimate the suppression masks for other subsequent frames. The suppression masks restrain the common salient information and thoroughly discover the remaining potential discriminative information of other subsequent frames. Extensive experiments on two benchmarks demonstrate the effectiveness and efficiency of our method, surpassing the state-of-the-art methods by a large margin.

The main contributions of this paper are as following: (1) We first regard I2V Re-ID as point-to-set matching problem, and propose a novel Temporal Complementarity-Guided Reinforcement Learning (TCRL) approach, towards achieving both efficiency and accuracy. (2) We formulate point-to-set matching procedure as Markov decision process, and train an agent to make sequential judgments on adaptively selecting suitable amount of frames from gallery videos by the guidance of a query image for recognizing pedestrians. (3) We design a sequential feature extraction module with complementary residual detectors to dynamically suppress common salient information and thoroughly mine potential complementary clues among these selected video frames for enhancing the capacity of frame-level features of pedestrians.

2. Related Work

Image and Video based Re-ID. Image and video based person Re-ID have been extensively studied in the past years [20, 32]. At the early stage, researchers pay more attention to design discriminative hand-crafted descriptors [6, 17, 29] and/or learn robust distance metric function [15, 23, 34, 45]. With the rise of deep learning technique, deep learning based methods have gained a great success and the performance is improved significantly on the widely-used image and video benchmarks. For example, Zhang *et al.* [42] proposed a deep graph model termed Heterogeneous Local Graph Attention Network, which models the inter-local relation and the intra-local relation in the completed local graph, simultaneously.

I2V Person Re-ID. Different from image and video-based person Re-ID, I2V person Re-ID [27] requires to learn heterogeneous features from image and video domains for matching pedestrians. Some works [30, 37, 47, 48] focus on projecting image and video embeddings into a shared feature space. For example, Zhu *et al.* [47] proposed a joint feature projection matrix and heterogeneous dictionary pair learning (PHDL) approach, which jointly learns an intra-video projection matrix and a pair of heterogeneous image and video dictionaries. Zhu *et al.* [37] proposed a cross-modality matching framework for I2V Re-ID, which adopts CNN and LSTM model for deep feature extraction and temporal information of video encoding, and further utilizes a neural network for similarity measure learning. Besides, some other methods [7, 25, 28] attempt to propagate the

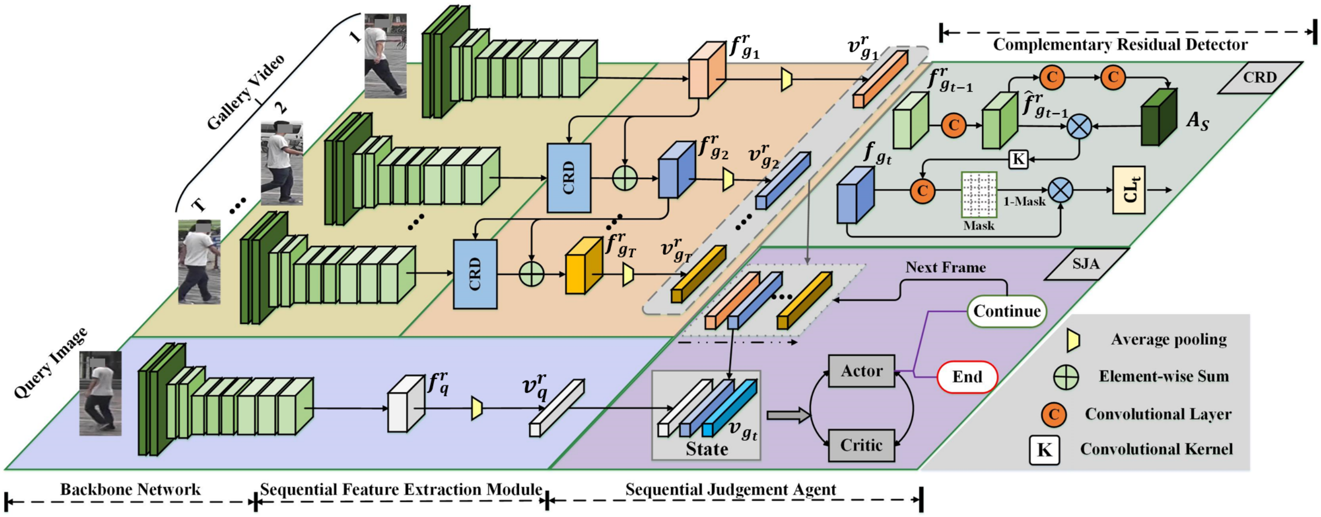


Figure 2. The overall architecture of the proposed TCRL. It consists of three components: a backbone network, a sequential feature extraction module (SFE) equipped with complementary residual detectors (CRD), and a sequential judgement agent (SJA).

temporal knowledge learned by video feature network to image feature network for solving information asymmetry problem. For example, Gu *et al.* [7] proposed a Temporal Knowledge Propagation framework to transfer temporal information from video embedding network to image one, and learn a shared feature. Shim *et al.* [28] proposed a Reciprocal Attention Discriminator along with two losses, which integrates asymmetric information of image-video pair by using non-local operation.

Reinforcement Learning. Reinforcement Learning (RL) aims at training an agent to learn the optimal policy by interacting with dynamic environment [16], which has been introduced in many computer vision tasks, *e.g.*, object detection [21,40] and visual tracking [12,26]. Recently, RL has been applied for image and video based person Re-ID to generate spatial or temporal attention [1, 22, 24, 38], but is not considered for I2V person Re-ID. For example, Chen *et al.* [2] proposed a recurrent 3-dimensional attentive reinforcement learning framework, which jointly attends to the salient body parts of person videos on both spatial and temporal dimensions. Li *et al.* [13] proposed a Deep Reinforcement Attention Learning (DREAL) framework, which facilitates visual recognition in a quality-aware manner, and employs recurrent critics that assess the attention action based on performance improvement.

3. Method

To effectively and efficiently learn discriminative pedestrian representations under point-to-set matching setting, we introduce a novel Temporal Complementarity-Guided Reinforcement Learning (TCRL) approach for image-to-video person re-identification. The overall detailed architecture of the proposed TCRL is illustrated in Figure 2. It mainly consists of three components: a backbone network , a sequential feature extraction module (SFE) equipped with complementary residual detectors (CRD) , and a sequential judgement agent (SJA).

3.1. Backbone Network

Given a gallery video sequence denoted as $\{\boldsymbol{I}_t\}_{t=1}^T$, the backbone network (the first four residual layers of ResNet-50 model [9]) is employed to extract the initial frame-level features $\{\boldsymbol{f}_{g_t}|\boldsymbol{f}_{g_t} \in \mathbb{R}^{H\times W\times C}\}_{t=1}^T$, where H, W and C are the height, weight and channel size of each feature $\boldsymbol{f}_{g_t}$, T indicates the total number of the video sequence and t is the index of the frame. Meanwhile, the backbone network also extracts the initial feature of the query image, which is denoted as $\boldsymbol{f}_q \in \mathbb{R}^{H\times W\times C}$.

3.2. Sequential Feature Extraction Module

In contrast to a single query image, a gallery video sequence contains rich spatial-temporal appearance clues, which are beneficial for learning robust feature representation. Existing video feature extractors are limited to perform the same operation on each video frame, leading to high redundant representations of different frames that only highlight nearly identical local regions [14]. Thus, we introduce a sequential feature extraction module equipped with complementary residual detectors to effectively mine complementary information from consecutive frames and learn more complete and informative frame-level features. Given the initial frame-level features of a gallery video $\{\boldsymbol{f}_{g_t}|\boldsymbol{f}_{g_t} \in \mathbb{R}^{H\times W\times C}\}_{t=1}^T$, the complementary residual detector is employed to surpass redundant salient regions have been activated in previous frames and explore complementary information within the current frame to enhance the frame-level representation. The overview of the complementary residual detector is shown in Figure 2. Specifically, in order to extract the reinforced frame-level feature, the detector is composed of three operations: salient region location, residual information excavation and complementary feature learning.

Salient Region Location. This operation aims to find the redundant salient features have been captured in previous frames. Specifically, we denote $\boldsymbol{f}^r_{g_{t-1}} \in \mathbb{R}^{H\times W\times C'}$ as the extracted reinforced frame-level feature of the $(t-1)$-th frame, containing the temporal complementary information of all previous $t-1$ frames. A 1×1 convolution layer is applied on $\boldsymbol{f}^r_{g_{t-1}}$ for projecting it into $\hat{\boldsymbol{f}}^r_{g_{t-1}} \in \mathbb{R}^{H\times W\times C}$, which has the same channel dimension with $\boldsymbol{f}_{g_t}$. To discover the most activated salient regions of previous $t-1$ frames, a simple yet effective multi-head attention mechanism (implemented by another two convolution layers) is employed to weight the importance of each spatial position and produce k^2 diverse attention maps. The multi-head attention mechanism is formulated as following:

$$\boldsymbol{A}_s = ReLU(g_{s_2}(BN(g_{s_1}(\hat{\boldsymbol{f}}^r_{g_{t-1}})))) \tag{1}$$

where g_{s_1} and g_{s_2} denotes two convolution layers, BN is batch normalization layer and $ReLU$ refers to a rectified linear unit layer. The output channels of g_{s_2} is k^2. The result $\boldsymbol{A}_s \in \mathbb{R}^{k^2\times H\times W}$ represents the varied salient regions of the feature map $\hat{\boldsymbol{f}}^r_{g_{t-1}}$. After that, the salient feature $\boldsymbol{S}_{t-1}$ of all these $t-1$ video frames are learned by $\boldsymbol{S}_{t-1} = \boldsymbol{A}_s\hat{\boldsymbol{f}}^r_{g_{t-1}} \in \mathbb{R}^{k^2\times C}$.

Residual Information Excavation. The operation suppresses the activated redundant salient regions before and mines remaining complementary regions of the latter frame for capturing residual complementary information. We firstly reshape $\boldsymbol{S}_{t-1}$ into $\boldsymbol{S}_{t-1} \in \mathbb{R}^{k\times k\times C\times 1}$, which is viewed as a salient convolutional kernel, with kernel size of $[k,k]$, input channels of C and output channels of 1. The salient convolutional kernel is then employed to perform a convolution operation on the initial frame-level feature $\boldsymbol{f}_{g_t}$ of the latter t-th frame. It is formulated as following:

$$\boldsymbol{M}_t = softmax(\boldsymbol{f}_{g_t} \otimes \boldsymbol{S}_{t-1}) \tag{2}$$

where $\otimes$ denotes the convolution operation, $softmax$ is applied on the $H\times W$ dimension for normalizaiton. $\boldsymbol{M}_t \in \mathbb{R}^{H\times W}$ is the suppression mask, which is an affinity matrix presenting high relevance values for the salient feature $\boldsymbol{S}_{t-1}$ captured in previous frames with respect to the t-th frame. It is worth noting that $\boldsymbol{M}_t$ in fact reflects the patch-wise similarity at every spatial location between the feature maps $\hat{\boldsymbol{f}}^r_{g_{t-1}}$ and $\boldsymbol{f}_{g_t}$. Moreover, the residual complementary information of the t-th frame is mined by

$$\boldsymbol{R}_t = (1-\boldsymbol{M}_t)\cdot \boldsymbol{f}_{g_t} \tag{3}$$

while the salient feature with high similarities is suppressed for the frame t.

Complementary Feature Learning. This operation utilizes the specific complementary learners (CL) to encode the residual complementary information and learn the reinforced frame-level features of the latter frame. Concretely, the complementary feature $\boldsymbol{f}^c_{g_t}$ is computed as following:

$$\boldsymbol{f}^c_{g_t} = CL_t(\boldsymbol{R}_t) \tag{4}$$

where CL_t is constructed by the last residual layer of ResNet-50 model for learning identity-related feature from the residual complementary information of frame t. It shares the same parameters in the first two residual blocks and has its specific parameters in the last block for different frames [10]. The different specific learners perform collaboratively to discover diverse complementary visual cues, towards generating integral characteristic of the identity. After that, the reinforced frame-level feature of the t-th frame is obtained as following:

$$\boldsymbol{f}^r_{g_t} = \boldsymbol{f}^c_{g_t} + \boldsymbol{f}^r_{g_{t-1}} \tag{5}$$

By recursively performing the complementary residual detector on all the frames sequentially, the corresponding reinforced frame-level features $\{\boldsymbol{f}^r_{g_t}\}^T_{t=1}$ of these frames of the gallery video are learned. Besides, the query image is also treated as a video sequence containing only one frame, and obtains its reinforced feature $\boldsymbol{f}^r_q$ in the same way. Such features $\{\boldsymbol{f}^r_{g_t}\}^T_{t=1}$ are finally applied with a global average pooling operation to produce the feature vectors $\{\boldsymbol{v}^r_{g_t}\}^T_{t=1}$.

3.3. Sequential Judgement Agent

Consecutive frames of a video usually contain vast redundant information, even noisy information caused by partial occlusion, cluttered background or inaccurate detection [14, 18, 30]. Therefore, the approach of directly using all frames to obtain the integral characteristic of a pedestrian is computationally inefficient, and gives rise to discrimination degradation of the learned representation. Considering that, we formulate I2V Re-ID problem as a Markov decision process (MDP), and utilize the sequential judgement agent to dynamically select suitable amount of video frames for pedestrian matching, which is illustrated in Figure 2. At each time step t, the agent takes the first t frames of the gallery video and the query image as a dynamic environment to observe the state $\boldsymbol{s}_t$, executes the action $\boldsymbol{a}_t$ from the learned experience, receives the reward $\boldsymbol{r}_t$, towards optimizing the policy. Then, if the current episode is not terminated, this agent will take one more frame of the gallery video and the query image to update the state $\boldsymbol{s}_{t+1}$, at the next time step $t+1$. The details of the state, action, reward and the architecture of the agent are elaborated below.

State. The state $\boldsymbol{s}_t$ at time step t in the episode consists of four components $\boldsymbol{s}_t = [\boldsymbol{v}^r_q, \boldsymbol{v}_{g_t}, \boldsymbol{v}^r_{g_t}, |\boldsymbol{v}^r_q - \boldsymbol{v}^r_{g_t}|]$, where $\boldsymbol{v}_q^r, \boldsymbol{v}_{g_t}^r \in \mathbb{R}^{C'}$ are the reinforced feature vectors of the query image and the t-th frame, respectively. $\boldsymbol{v}_{g_t} \in \mathbb{R}^{C'}$ is the initial feature vector of the t-th frame learned from the backbone network with the specific complementary learner. These four components are essential and complementary, because $\boldsymbol{v}_q^r$ and $\boldsymbol{v}_{g_t}$ provide visual content of current time, $\boldsymbol{v}_{g_t}^r$ contains the historical complementary information of all previous frames and $|\boldsymbol{v}_q^r - \boldsymbol{v}_{g_t}^r|$ represents the feature affinity of the query image and the gallery video.

Agent. We adopt deep deterministic policy gradient (DDPG) [16] to construct the sequential judgement agent, which consists of four parts, *i.e.*, an actor, a critic, a target actor and a target critic. These four parts are all implemented with three fully connected layers. The last fully connected layer of the actor and target actor is followed by a Sigmoid function to maintain the value of the action $\boldsymbol{a}_t$ lies between 0 and 1.

Action. The agent defines two types of actions: *continue* or *end*. The former action indicates the agent requires one more frame from the gallery video for distinguishing different pedestrians. The latter one indicates immediately terminating the current episode, which means the agent has accumulated enough temporal complementary information from the limited number of frames to make the judgement of distinguishing the identities, avoiding unnecessary computation. Specifically, when the value of the action $\boldsymbol{a}_t < 0.5$, the agent requires to explore next frame, and when the value of the action $\boldsymbol{a}_t \geq 0.5$, the current episode should be terminated. Besides, when the agent comes to the last frame of the gallery video, it is forced to terminate.

Reward. The reward reflects the value of the action executed by the agent with regard to the state. We define the reward at time step t as following:

$$\boldsymbol{r}_t = \begin{cases} r_0 & \text{, if stimulate} \\ -r_0 & \text{, if punishment} \\ \mathrm{Tanh}(1 - \frac{t}{\beta}) \cdot r_0 & \text{, otherwise} \end{cases} \tag{6}$$

At each time step t, we apply L_2 normalization on the channel dimension of the feature vectors $\boldsymbol{v}_q^r$ and $\boldsymbol{v}_{g_t}^r$, and measure the similarity score as $\boldsymbol{z}_t = \boldsymbol{v}_q^r \cdot {\boldsymbol{v}_{g_t}^r}^T$. If the actor requires the next frame to assist pedestrian recognition ($\boldsymbol{a}_t < 0.5$), we compute the reward ($\boldsymbol{r}_t = \mathrm{Tanh}(1 - \frac{t}{\beta}) \cdot r_0$) based on frame index t, which encourages the agent to gather more temporal complementary information for enhancing feature representation when t is tiny at the beginning, while punishes it for exploring excess frames when t is large to ensure high efficiency. The similarity score between the current frame t and the query image is recorded in buffer $\mathcal{Z}$. Besides, if the actor terminates the current episode ($a_t \geq 0.5$), we stimulate the agent ($\boldsymbol{r}_t = r_0$) in two situation: 1) if $\boldsymbol{z}_t > max(\mathcal{Z})$ and the query image has the same identify with the gallery video; 2) if $\boldsymbol{z}_t < min(\mathcal{Z})$ and the query image has different identity with the gallery video. And we give the agent a punishment ($\boldsymbol{r}_t = -r_0$) in the other situation.

3.4. Model Optimizing

Triplet loss and identification loss [36, 42] are widely used in task of person re-identification. We adopt the frame-wise triplet loss with hard mining strategy and frame-wise identification loss with label smoothing regularization to train the sequential feature extraction module and optimize the feature vectors $\boldsymbol{v}_q^r$ and $\{\boldsymbol{v}_{g_t}^r\}_{t=1}^{T'}$ (T' is the index of the selected last frame of the gallery video by the sequential judgement agent.) Triplet loss and identification loss are denoted as $\mathcal{L}_{\text{tri}}$ and $\mathcal{L}_{\text{ide}}$, respectively. The total loss for SFE is computed as following:

$$\mathcal{L}_{\text{sfe}} = \mathcal{L}_{\text{tri}} + \mathcal{L}_{\text{ide}} \tag{7}$$

In order to optimize the actor and critic of the agent, we first random sample a batch of data $\mathcal{B} = \{(\boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{r}_t, \boldsymbol{s}_{t+1}, \boldsymbol{d}_t)\}$ (where $\boldsymbol{d}_t$ indicates whether the current episode is terminal) in the replay buffer $\mathcal{D}$, and then compute the target long-term reward $\boldsymbol{R}_t$ [16] as following:

$$\boldsymbol{R}_t = \boldsymbol{r}_t + \gamma(1 - \boldsymbol{d}_t)\mathcal{C}_{target}(\boldsymbol{s}_{t+1}, \mathcal{A}_{target}(\boldsymbol{s}_{t+1})) \tag{8}$$

Where $\mathcal{C}_{target}$ is the target critic, $\mathcal{A}_{target}$ is the target actor, and $\gamma \in [0, 1]$ is a discount factor. The critic network $\mathcal{C}$ of the agent is optimized with mean squared error (MSE) loss:

$$L_{\text{crt}} = \underset{(\boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{r}_t, \boldsymbol{s}_{t+1}, \boldsymbol{d}_t) \sim \mathcal{D}}{\mathrm{E}}[(\mathcal{C}(\boldsymbol{s}_t, \boldsymbol{a}_t) - \boldsymbol{R}_t)^2] \tag{9}$$

The actor $\mathcal{A}$ of the agent is optimized by performing gradient descent to solve:

$$L_{\text{act}} = \underset{\boldsymbol{s}_t \sim \mathcal{D}}{-\mathrm{E}}[\mathcal{C}(\boldsymbol{s}_t, \mathcal{A}(\boldsymbol{s}_t))] \tag{10}$$

The target critic and actor are updated by performing exponential moving average policy on the critic and actor.

4. Experiments

4.1. Experiment Settings

Datasets. MARS [44] is one of the largest video benchmark, which consists of 20,715 video sequences of 1,261 identities. The training set contain 625 identity, each identity has 13.2 video sequences on average and each video sequence has 59.5 frames on average. During inference, the query set and gallery set has 636 identities. The first frame of each query video is employed as query image to perform I2V Re-ID, following [7]. **iLIDS-VID** [31] is a small-scale video benchmark, which contains 600 video sequences of 300 identities. The length of each video sequence changes from 22 to 192 frames, with an average of 71 frames. The

Table 1. Performance comparison to the state-of-the-art methods on MARS dataset.

Method	Rank-1	Rank-5	Rank-10	mAP
MPHDL [48]	39.1	53.8	63.3	-
P2SNet [30]	55.3	72.9	78.7	-
TMSL [7]	56.5	70.6	-	-
XQDA [4]	67.2	81.9	86.1	54.9
TKP [7]	75.6	87.6	90.9	65.1
DSA [41]	78.3	88.9	91.4	68.7
STE-NVAN [18]	80.3	-	-	68.8
SAA-CMIL [27]	81.3	91.7	93.8	72.6
READ [28]	81.5	91.2	93.3	69.9
ResVKD [25]	83.9	93.2	-	77.3
TCRL	**86.0**	92.5	**94.2**	**80.1**

whole dataset is randomly split into a training set, a query set, and a gallery set. The training set has 150 identities, and the query set and gallery set share the rest 150 identities. During inference, the first frames of all videos captured by the first camera are employed to perform I2V person Re-ID, consistent with the methods [30, 37].

Implementation Details. We adopt ResNet-50 model [9] pre-trained on ImageNet in our TCRL. The last stride of ResNet-50 model is set to 1. At the training stage, each min-batch contains 4 identities and each identity has 4 input video clips. We randomly sample $T = 6$ frames from video sequences as input clip. All frames are resized into 256×128 pixels. Random flipping strategy is used for data augmentation. The Adam optimizer is adopted to optimize the sequential feature extraction module with the initial learning rate of $3e^{-4}$ and the weight decay of $5e^{-4}$ for 150 epochs. The learning rate is decayed by 0.1 every 40 epochs. After that, we randomly choose positive or negative query image and gallery video pairs with ratio $1 : 1$ to train the sequential judgement agent and the sequential feature extraction module. We train the two components for 50 epoch with Adam optimizer. The learning rates of the actor and critic for the agent are set to be $1e^{-4}$ and $1e^{-5}$, respectively. The kernel size k of CRD is set to 3. r_0 and β in Eq. 6 are set to 1.0 and 1.9, respectively. The discount factor γ is 0.99. All the experiments on the two benchmarks follow the same settings described above. During inference, we follow [18] to sample the first frames from T equally-divided chunks for producing the input video clips.

Evaluation Metrics. Two standard metrics are adopted to evaluate the performance of I2V Re-ID algorithms, that are Cumulative Matching Characteristic at Rank-1, Rank-5, Rank-10, and mean average precision (mAP).

Table 2. Performance comparison to the state-of-the-art methods on iLIDS-VID dataset.

Method	Rank-1	Rank-5	Rank-10
PSDML [46]	13.5	33.8	45.6
LERM [11]	15.3	37.1	49.7
PHDL [47]	28.2	50.4	65.9
MPHDL [48]	32.6	55.8	69.3
TMSL [7]	39.5	66.9	79.6
P2SNet [30]	40.0	68.5	78.1
CME [33]	40.1	67.2	79.7
TKP [7]	54.6	79.4	86.9
SAA-CMIL [27]	54.7	78.0	87.3
TCRL	**77.3**	**94.7**	**96.7**

4.2. Comparison to State-of-the-Art Methods

Result on MARS. In Table 1, we compare the performance of the proposed TCRL with 10 state-of-the-art approaches on MARS dataset. We can observe that TCRL outperforms all these methods by a large margin, particularly surpassing the second best method ResVKD [25] by 2.1% Rank-1 accuracy and 2.8% mAP. The comparison clearly validates the effectiveness and superiority of TCRL. The main reasons for the obvious improvement are: 1) the sequential judgement agent selects suitable number of frames from gallery video clips, which reduces the influence of noisy information from redundant frames; 2) the sequential feature extraction module enhances the capacity of frame-level feature by thoroughly mining potential complementary clues among these selected frames. More importantly, compare with all these methods that regard I2V person Re-ID as a cross-modality task and attempt to directly project image and video features into the common embedding space or propagate the temporal knowledge, TCRL treats I2V Re-ID as a point-to-set matching task, which can achieve both efficiency and accuracy.

Result on iLIDS-VID. We compare TCRL with 9 state-of-the-art approaches on iLIDS-VID dataset in Table 2. The proposed TCRL reaches the best re-identification performance of 77.3% Rank-1 accuracy, outperforming all existing methods by a large margin. It improves the second best method SAA-CMIL [27] by 22.6% Rank-1 accuracy. The comparison demonstrates the powerful capability and applicability of the proposed TCRL on the relatively small person Re-ID benchmark.

4.3. Ablation Study

Effectiveness of Components. To verify the impact of each component within TCRL, we report the results of ab-

Table 3. Evaluation of the effectiveness of each component of TCRL on MARS dataset.

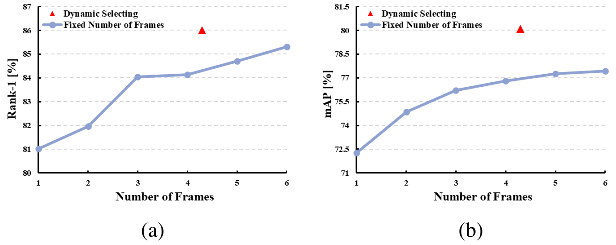
Model	Frames	Rank-1	Rank-5	Rank-10	mAP
Basel	6.0	82.8	92.0	93.5	74.4
Basel+SFE	6.0	85.3	92.9	94.0	78.8
Basel+SFE+SJA	4.2	86.0	92.5	94.2	80.1

Table 4. Analysis on the influence of parameter β for SJA on MARS dataset.

β	Frames	Rank-1	Rank-5	Rank-10	mAP
1.8	3.2	84.3	92.1	93.7	77.8
1.9	4.2	86.0	92.5	94.2	80.1
2.0	5.5	85.2	93.0	94.1	78.9

lation study in Table 3. Basel denotes only using ResNet-50 model to extract the image-level and video-level representation. Basel+SFE represents that we utilize the backbone network and SFE equipped with CRD to extract the reinforced query and frame-level features by utilizing all frames. Basel+SFE+SJA refers that we employ the whole TCRL to dynamically select the suitable amount of frames from gallery videos to further improve the capacity of the reinforced features, as well as enhance the model efficiency. Compared with Basel, Basel+SFE improves the performance on Rank-1 accuracy and mAP by 2.5% and 4.4%, respectively. It demonstrates the effectiveness of SFE to mine the complementary residual clues and suppress redundant salient regions among video frames for learning reinforced frame-level features. Moreover, Basel+SFE+SJA surpasses Basel+SFE by 0.7% Rank-1 accuracy and 1.3% mAP. It indicates the ability of the sequential judgement agent to select suitable number of necessary frames from gallery videos, which further enhances the frame-level representation by reducing the interference from low quality frames, and improves the efficiency by only using 4.2 frames on average for all gallery videos during testing.

Analysis of Sequential Judgement Agent. In Figure 3, we verify the effect of dynamically selecting suitable amount of frames by SJA, and compare Rank-1 accuracy and mAP of it with using fixed number of frames. The blue dots indicate the performance of using fixed length of video sequences. The red triangle indicates the performance of dynamically selecting suitable amount of frames. We can observe that the agent only uses 4.2 frames on average to make the judgment on distinguishing different identities, and achieves the best performance. It surpasses the performance of using fixed number of frames with varying from 1 to 6. The comparison demonstrates that part of all video frames can provide sufficient discriminative information for recognizing pedestrians, and fixedly using more frames brings about lower efficiency while fixedly using less frames brings about performance degradation. Thus, it is significance to adaptively select suitable amount of frames for each gallery video according to query image, towards balancing efficiency and accuracy. To further investigate the decision of the agent for each gallery video, we record the number of used frames for all gallery videos. The result is illustrated in Figure 4. It indicates the ability of the sequential judgement agent to dynamically choose applicable number of imperative frames for gallery videos, and enhance efficiency.



(a) (b)

Figure 3. Evaluation of the influence of dynamically using suitable amount of frames and using fixed number of frames.

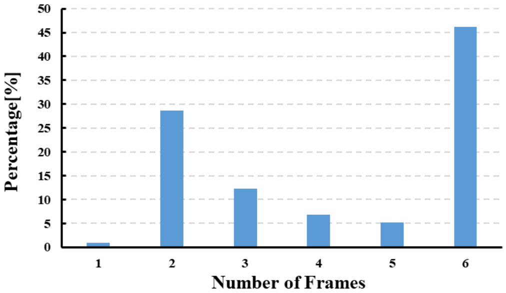


Figure 4. Statistics of the number of used frames for different gallery videos during testing on MARS dataset.

Moreover, in Table 4, we also investigate the influence of the key parameter β on the sequential judgement agent. When β is small, the reward quickly becomes negative and small in the subsequent time steps, which enforces the agent to stop searching more useful frames early. On the contrary, when β is large, the reward is large or even maintains positive in the subsequent time steps, which encourages the agent to gather excess useless information. Different settings of β make significant influence on the performance of the agent. From the comparison, we observe that the best Rank-1 accuracy and mAP are obtained at $\beta = 1.9$.

Analysis of Sequential Feature Extraction Module. We conduct experiments to analysis the influence of the

Table 5. Evaluation of the sequential feature extraction module with different settings on MARS dataset.

Model	Rank-1	Rank-5	Rank-10	mAP
SFE (T=5)	84.3	92.1	93.8	77.0
SFE (T=6)	85.3	92.9	94.0	78.8
SFE (T=7)	85.0	92.5	94.1	77.6
SFE (k=1)	84.4	92.7	93.2	77.3
SFE (k=3)	85.3	92.9	94.0	78.8
SFE (k=5)	84.7	92.5	93.4	78.2

Table 6. Computational complexity comparison with the state-of-the-art methods on MARS dataset.

Model	Number of Params	Number of Frames
TKP [7]	54.4M	7.4
ResVKD [25]	47.0M	8.0
TCRL	35.0M	4.2

length of video sequences T and the size of the salient convolutional kernel k for SFE. In Table 5, we can observe that the sequential feature extraction module achieves best result with $T = 6$. This implies when $T = 5$, the video sequences can not provide enough complementary information for SFE. When $T = 7$, the gain diminish which is brought by introducing more noise information from more frames for SFE. In addition, we find that the performance of $k = 3$ is superior to other settings. We consider that the salient convolutional kernel is not able to filter salient regions when $k < 3$. And when k is too big, the kernel covers both the salient and sub-salient regions, which may lead to the collapse of the performance.

Computational Complexity. To verify the high efficiency of our TCRL, we report the total amount of parameters and the number of used frames on average of all gallery videos for TCRL and the state-of-the-art methods, in Table 6. It can be seen that the computation cost of TCRL is far less than ResVKD [25] and TKP [7]. Besides, compared with them, TCRL requires fewer number of frames for gallery videos to obtain better re-identification performance. Such indicates the efficiency and effectiveness of our TCRL. *Note that due to GPU limitation, fixed length of $T = 6$ of video clips are generated as input from gallery videos. If given video clips with much more frames, TCRL can improve the efficiency more significantly.*

Visualization Results. Figure 5 gives the visualization results of the learned feature maps of the selected frames for two video sequences by TCRL. We can see that TCRL is able to capture the salient region for the first frames. For subsequent frames, TCRL suppresses the redundant salient region, and focuses on a broad complementary region with informative clues to cover nearly whole foreground and learn more discriminative frame-level feature. Such verifies the powerful capacity of TCRL to thoroughly mine temporal complementary information across video frames for enhancing feature representation.

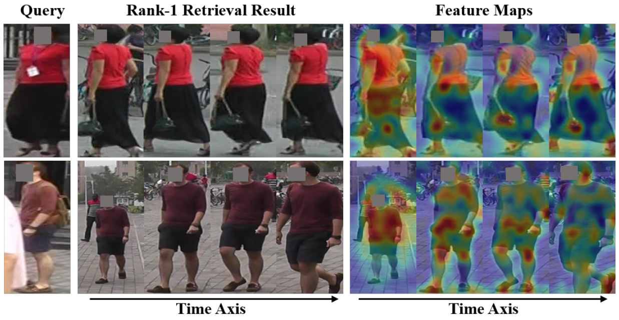


Figure 5. Visualization results of the learned feature maps of the selected frames for two video sequences on MARS dataset.

5. Conclusion

In this work, we first regard I2V Re-ID as point-to-set matching problem, and propose a novel Temporal Complementarity-Guided Reinforcement Learning (TCRL) approach, towards achieving both efficiency and accuracy. TCRL formulates point-to-set matching procedure as Markov decision process, and trains a sequential judgement agent to measure the uncertainty between the query image and all historical frames of gallery videos at each time step, and learn the optimal policy to adaptively select suitable amount of frames to make the decision on whether the model has collected enough evidence for identifying the same pedestrian or distinguishing different pedestrians. Besides, TCRL maintains a sequential feature extraction module with complementary residual detectors to dynamically suppress redundant salient regions activated in previous frames and thoroughly mine new and potential complementary clues among subsequent frames for enhancing frame-level representation. Extensive experiments verify the superiority of the proposed TCRL.

Acknowledgment

This work was supported by the National Key R&D Program of China under Grant 2020AAA0105702, National Natural Science Foundation of China (NSFC) under Grant U19B2038 and 62106245, the University Synergy Innovation Program of Anhui Province under Grants GXXT-2019-025, and the Fundamental Research Funds for the Central Universities under Grant WK2100000021.

References

[1] Guangyi Chen, Chunze Lin, Liangliang Ren, Jiwen Lu, and Jie Zhou. Self-critical attention learning for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9637–9646, 2019. 3

[2] Guangyi Chen, Jiwen Lu, Ming Yang, and Jie Zhou. Learning recurrent 3d attention for video-based person re-identification. *IEEE Transactions on Image Processing*, 29:6963–6976, 2020. 3

[3] Guangyi Chen, Yongming Rao, Jiwen Lu, and Jie Zhou. Temporal coherence or temporal motion: Which is more critical for video-based person re-identification? In *Proceedings of the European Conference on Computer Vision*, pages 660–676. Springer, 2020. 1

[4] Husheng Dong, Ping Lu, Shan Zhong, Chunping Liu, Yi Ji, and Shengrong Gong. Person re-identification by enhanced local maximal occurrence representation and generalized similarity metric learning. *Neurocomputing*, 307:25–37, 2018. 6

[5] Chanho Eom, Geon Lee, Junghyup Lee, and Bumsub Ham. Video-based person re-identification with spatial and temporal memory networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 12036–12045, 2021. 1

[6] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. Person re-identification by symmetry-driven accumulation of local features. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2360–2367. IEEE, 2010. 2

[7] Xinqian Gu, Bingpeng Ma, Hong Chang, Shiguang Shan, and Xilin Chen. Temporal knowledge propagation for image-to-video person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9647–9656, 2019. 2, 3, 5, 6, 8

[8] Jianyuan Guo, Yuhui Yuan, Lang Huang, Chao Zhang, Jin-Ge Yao, and Kai Han. Beyond human parts: Dual part-aligned representations for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3642–3651, 2019. 1

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 3, 6

[10] Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Temporal complementary learning for video person re-identification. In *Proceedings of the European Conference on Computer Vision*, pages 388–405. Springer, 2020. 4

[11] Zhiwu Huang, Ruiping Wang, Shiguang Shan, and Xilin Chen. Learning euclidean-to-riemannian metric for point-to-set classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1677–1684, 2014. 6

[12] Mingxin Jiang, Tao Hai, Zhigeng Pan, Haiyan Wang, Yinjie Jia, and Chao Deng. Multi-agent deep reinforcement learning for multi-object tracker. *IEEE Access*, 7:32400–32407, 2019. 3

[13] Duo Li and Qifeng Chen. Deep reinforced attention learning for quality-aware visual recognition. In *Proceedings of the European Conference on Computer Vision*, pages 493–509. Springer, 2020. 3

[14] Shuang Li, Slawomir Bak, Peter Carr, and Xiaogang Wang. Diversity regularized spatiotemporal attention for video-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 369–378, 2018. 3, 4

[15] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2197–2206, 2015. 2

[16] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015. 3, 5

[17] Chunxiao Liu, Shaogang Gong, Chen Change Loy, and Xinggang Lin. Person re-identification: What features are important? In *Proceedings of the European Conference on Computer Vision*, pages 391–401. Springer, 2012. 2

[18] Chih-Ting Liu, Chih-Wei Wu, Yu-Chiang Frank Wang, and Shao-Yi Chien. Spatially and temporally efficient non-local attention network for video-based person re-identification. *arXiv preprint arXiv:1908.01683*, 2019. 4, 6

[19] Jiawei Liu, Zheng-Jun Zha, Di Chen, Richang Hong, and Meng Wang. Adaptive transfer network for cross-domain person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2019. 1

[20] Jiawei Liu, Zheng-Jun Zha, Xuejin Chen, Zilei Wang, and Yongdong Zhang. Dense 3d-convolutional neural network for person re-identification in videos. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 15(1s):1–19, 2019. 2

[21] Lijie Liu, Chufan Wu, Jiwen Lu, Lingxi Xie, Jie Zhou, and Qi Tian. Reinforced axial refinement network for monocular 3d object detection. In *Proceedings of the European Conference on Computer Vision*, pages 540–556. Springer, 2020. 3

[22] Zimo Liu, Jingya Wang, Shaogang Gong, Huchuan Lu, and Dacheng Tao. Deep reinforcement active learning for human-in-the-loop person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6122–6131, 2019. 3

[23] Lianyang Ma, Xiaokang Yang, and Dacheng Tao. Person re-identification over camera networks using multi-task distance metric learning. *IEEE Transactions on Image Processing*, 23(8):3656–3670, 2014. 2

[24] Deqiang Ouyang, Jie Shao, Yonghui Zhang, Yang Yang, and Heng Tao Shen. Video-based person re-identification via self-paced learning and deep reinforcement learning framework. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1562–1570, 2018. 3

[25] Angelo Porrello, Luca Bergamini, and Simone Calderara. Robust re-identification by multiple views knowledge distillation. In *Proceedings of the European Conference on Computer Vision*, pages 93–110. Springer, 2020. 2, 6, 8

[26] Liangliang Ren, Jiwen Lu, Zifeng Wang, Qi Tian, and Jie Zhou. Collaborative deep reinforcement learning for multi-object tracking. In *Proceedings of the European Conference on Computer Vision*, pages 586–602, 2018. 3

[27] Wei Shi, Hong Liu, and Mengyuan Liu. Image-to-video person re-identification using three-dimensional semantic appearance alignment and cross-modal interactive learning. *Pattern Recognition*, 122:108314, 2021. 2, 6

[28] Minho Shim, Hsuan-I Ho, Jinhyung Kim, and Dongyoon Wee. Read: Reciprocal attention discriminator for image-to-video re-identification. In *Proceedings of the European Conference on Computer Vision*, pages 335–350. Springer, 2020. 2, 3, 6

[29] Rahul Rama Varior, Gang Wang, Jiwen Lu, and Ting Liu. Learning invariant color features for person reidentification. *IEEE Transactions on Image Processing*, 25(7):3395–3410, 2016. 2

[30] Guangcong Wang, Jianhuang Lai, and Xiaohua Xie. P2snet: Can an image match a video for person re-identification in an end-to-end way? *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10):2777–2787, 2017. 2, 4, 6

[31] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by video ranking. In *Proceedings of the European Conference on Computer Vision*, pages 688–703. Springer, 2014. 5

[32] Bryan Ning Xia, Yuan Gong, Yizhe Zhang, and Christian Poellabauer. Second-order non-local attention networks for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3760–3769, 2019. 2

[33] Zhongwei Xie, Lin Li, Xian Zhong, Luo Zhong, and Jianwen Xiang. Image-to-video person re-identification with cross-modal embeddings. *Pattern Recognition Letters*, 133:70–76, 2020. 2, 6

[34] Fei Xiong, Mengran Gou, Octavia Camps, and Mario Sznaier. Person re-identification using kernel-based metric learning methods. In *Proceedings of the European Conference on Computer Vision*, pages 1–16. Springer, 2014. 2

[35] Jinrui Yang, Jiawei Zhang, Fufu Yu, Xinyang Jiang, Mengdan Zhang, Xing Sun, Ying-Cong Chen, and Wei-Shi Zheng. Learning to know where to see: A visibility-aware approach for occluded person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 11885–11894, 2021. 1

[36] Jinrui Yang, Wei-Shi Zheng, Qize Yang, Ying-Cong Chen, and Qi Tian. Spatial-temporal graph convolutional network for video-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3289–3299, 2020. 1, 5

[37] Dongyu Zhang, Wenxi Wu, Hui Cheng, Ruimao Zhang, Zhenjiang Dong, and Zhaoquan Cai. Image-to-video person re-identification with temporally memorized similarity learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10):2622–2632, 2017. 2, 6

[38] Jianfu Zhang, Naiyan Wang, and Liqing Zhang. Multi-shot pedestrian re-identification via sequential decision making. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6781–6789, 2018. 3

[39] Wei Zhang, Shengnan Hu, Kan Liu, and Zhengjun Zha. Learning compact appearance representation for video-based person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(8):2442–2452, 2019. 1

[40] Wei Zhang, Ran Song, Yibin Li, et al. Online decision based visual tracking via reinforcement learning. *Proceedings of the Advances in Neural Information Processing Systems*, 33, 2020. 3

[41] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Densely semantically aligned person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2019. 6

[42] Zhong Zhang, Haijia Zhang, and Shuang Liu. Person re-identification using heterogeneous local graph attention networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12136–12145, 2021. 1, 2, 5

[43] Kecheng Zheng, Wu Liu, Lingxiao He, Tao Mei, Jiebo Luo, and Zheng-Jun Zha. Group-aware label transfer for domain adaptive person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5310–5319, 2021. 1

[44] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. Mars: A video benchmark for large-scale person re-identification. In *Proceedings of the European Conference on Computer Vision*, pages 868–884. Springer, 2016. 5

[45] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Reidentification by relative distance comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(3):653–668, 2012. 2

[46] Pengfei Zhu, Lei Zhang, Wangmeng Zuo, and David Zhang. From point to set: Extend the learning of distance metrics. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2664–2671, 2013. 6

[47] Xiaoke Zhu, Xiao-Yuan Jing, Fei Wu, Yunhong Wang, Wangmeng Zuo, and Wei-Shi Zheng. Learning heterogeneous dictionary pair with feature projection matrix for pedestrian video retrieval via single query image. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017. 2, 6

[48] Xiaoke Zhu, Xiao-Yuan Jing, Xinge You, Wangmeng Zuo, Shiguang Shan, and Wei-Shi Zheng. Image to video person re-identification by learning heterogeneous dictionary pair with feature projection matrix. *IEEE Transactions on Information Forensics and Security*, 13(3):717–732, 2017. 2, 6