Pattern Recognition 122 (2022) 108314

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Image-to-video person re-identification using three-dimensional semantic appearance alignment and cross-modal interactive learning

Wei Shi[a], Hong Liu[a,*], Mengyuan Liu[b]

[a] Key Laboratory of Machine Perception, Peking University, Shenzhen Graduate School, Beijing 100871, China
[b] School of Intelligent Systems Engineering, Sun Yat-sen University, Guangzhou 510275, China

ARTICLE INFO

Article history:
Received 19 November 2020
Revised 20 June 2021
Accepted 9 September 2021
Available online 20 September 2021

2010 MSC:
68T10
68U10

Keywords:
Person re-identification
Cross-modal learning
Appearance alignment

ABSTRACT

Image-to-video person re-identification (I2V ReID), which aims to retrieve human targets between image-based queries and video-based galleries, has recently become a new research focus. However, the appearance misalignment and modality misalignment in both images and videos caused by pose variations, camera views, misdetections, and different data types, make I2V ReID still challenging. To this end, we propose a deep I2V ReID pipeline based on three-dimensional semantic appearance alignment (3D-SAA) and cross-modal interactive learning (CMIL) to address the aforementioned two challenges. Specifically, in the 3D-SAA module, the aligned local appearance images extracted by dense 3D human appearance estimation are in conjunction with global image and video embedding streams to learn more fine-grained identity features. The aligned local appearance images are further semantically aggregated by the proposed multi-branch aggregation network to weaken the negligible body parts. Moreover, to overcome the influence of modality misalignment, a CMIL module enables the communication between global image and video streams by interactively propagating the temporal information in videos to the channels of image feature maps. Extensive experiments on challenging MARS, DukeMTMC-VideoReID and iLIDS-VID datasets, show the superiority of our approach.

1. Introduction

Person re-identification (ReID) as a key component in multi-camera multi-target tracking, plays an important role in intelligent surveillance [1–3] and video analysis [4]. In recent years, abundant approaches have been proposed to address person ReID under the same modality, like image-based person ReID [5–8] and video-based person ReID [9–13]. Despite the best efforts of many researchers, existing person ReID methods under the same modality still can not be well applied to person ReID under different modalities, such as the identification between the image-based query and video-based gallery.

Image-to-video person re-identification (I2V ReID), is proposed to address problems mentioned above. In real scenarios, if only a single photo (query) of the criminal is captured, it is challenging to search the criminal among lots of surveillance videos (gallery). The main reason is the uncertainty of data due to the appearance and modality misalignment, as shown in Fig. 1. These two misalignment problems increase the intra-class variations, and make the existing single modality based ReID methods unable to be directly applied to I2V ReID task. It is well-known that the appearance information [5] and motion information [14] are crucial cues to identify persons in real surveillance scenarios. Therefore, how to integrate the rich temporal motion information in videos and spatial appearance information in images has become the focus of I2V ReID.

An intuitive solution to I2V ReID is to map both images and videos into the same compact feature space for the subsequent matching. Existing approaches generally utilize a Convolutional Neural Network (CNN) [16,17] based model to represent the appearance features of images, and a Long Short-Term Memory (LSTM) [16] / 3DCNN [18] / Non-Local CNN [19] model to learn spatio-temporal features from videos. Afterwards, a well-designed distance metric function is used to measure the difference among different identities or modalities. These solutions promote the improvement of I2V ReID, but still cannot completely solve the appearance and modality misalignment problems.

More specially, some researchers focus on extracting body part features based on human 2D joints to enhance the discrimination ability of learned features [20]. That is a good choice to introduce local features to I2V ReID. However, the 2D joints only roughly reflect the 2D center coordinates of key human body re-

* Corresponding author
E-mail addresses: pkusw@pku.edu.cn (W. Shi), hongliu@pku.edu.cn (H. Liu), nkliuyifang@gmail.com (M. Liu).

https://doi.org/10.1016/j.patcog.2021.108314
0031-3203/

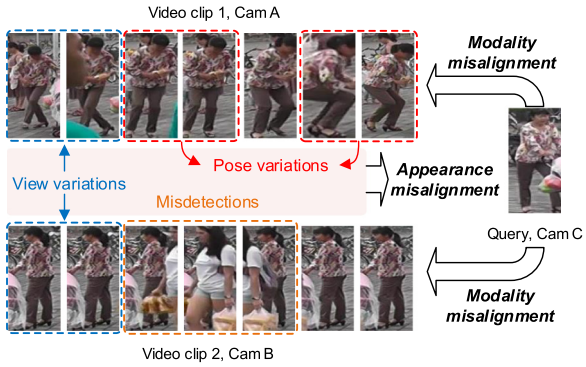

Fig. 1. Illustration of appearance and modality misalignment challenges in I2V ReID task. The appearance cues in the same video clip are sensitive to human pose variations, camera view variations and misdetections. We call this problem that affects the appearance of human as appearance misalignment. Moreover, the gallery videos in I2V ReID contain rich temporal motion information and appearance information, while the query images contain only appearance information. This kind of modality gap between images and videos in I2V ReID is called as modality misalignment. All the examples are selected from the MARS dataset [15].

gions, which is not helpful to obtain detailed 3D surface information of the human body, especially in the case of similar poses [21]. Inspired by the work [22], the more dense human 3D surface information is crucial to learn discriminative semantic features against appearance misalignment. Others dedicate to further minimize the distance between image and video modalities by introducing proxy text space [23], unsupervised domain adaption [18], temporal knowledge propagation [19], or self-attention mechanism [24]. Although these attempts are effective, they either need to introduce additional feature space, or force the two modalities to be unified into one. The image and video modalities in I2V ReID should be able to interactively communicate with each other while retaining the unique property of each modality.

In this paper, we aim to develop a more generalizable I2V ReID pipeline against the challenges mentioned above. The proposed I2V ReID pipeline contains two key components, three-dimensional semantic appearance alignment (3D-SAA) and cross-modal interactive learning (CMIL) module. Given the query images and gallery video clips, all query images and gallery frames in videos are fed into dense 3D human appearance estimation part in the 3D-SAA module to extract aligned local appearance images. Owe to the unified 3D human parametric model, the body parts in all query images and gallery frames are implicitly aligned. The extracted aligned local appearance images are further weighted and semantically aggregated to highlight more distinguished foreground texture parts by multi-branch aggregation network (MBAN) in the 3D-SAA module. The raw video clips, query images, and aligned local appearance images are simultaneously utilized to learn deep identity feature embeddings. Due to the diverse characteristics of different modalities, the image features contain rich semantics of appearances of target person, while the video features contain abundant temporal information. To this end, the CMIL module is proposed to interactively propagate modality-specific knowledge between two modalities. With the help of an interactive similarity comparison mechanism, the relation between image and video modalities is constructed, and is integrated into the channels of image features for the joint learning of two modalities.

Generally, our contributions are three-fold:

- A deep I2V ReID pipeline with two key components, three-dimensional semantic appearance alignment (3D-SAA) and cross-modal interactive learning (CMIL), is proposed to learn fine-grained and temporal invariant features, which achieves superior performances than the compared baseline method on MARS, DukeMTMC-VideoReID and iLIDS-VID datasets.
- To address the problem of appearance misalignment, a 3D-SAA module with proposed multi-branch aggregation network (MBAN) is designed to semantically align different body parts of human in the dense 3D human surface space, weaken the influence of negligible body parts, and aggregate different body parts into a unified appearance feature embedding.
- To address the problem of modality misalignment, a CMIL module is developed to construct the relation between two modalities with an interactive similarity comparison mechanism, and integrate the relation into the channels of image features for the interactive learning of two modalities.

The remainder of this paper is organized as follows. The related work is reviewed in Section 2. Section 3 clarifies the definition of I2V ReID and introduces the proposed pipeline with 3D-SAA and CMIL modules in detail. Experimental results and analysis are shown in Section 4, and Section 5 concludes this work.

2. Related work

2.1. Image-to-video person re-identification

Compared with single modality based person ReID tasks, like image-based person ReID [1,3] and video-based person ReID [9–13], I2V ReID indeed belongs to a cross-modal retrieval task between images and videos. Zhu et al. [25] and Li et al. [26] adopted heterogeneous dictionary pair learning and salient region clustering approach to tackle this task in a traditional manner, respectively. With the advance of deep learning, Zhang et al. [17] and Wang et al. [16] mapped raw images and videos to the learned heterogeneous deep feature space, and supervised the learning process by deep distance metrics. Specifically, Zhang et al. [17] utilized CNN to extract the features of images and frames of videos, and integrated frame-level features into a video-level feature with LSTM model. Wang et al. [16] also mapped images and frames of videos by CNN, and they designed a k-nearest neighbor triplet loss to constrain the relations between image-level features and frame-level features across different identities. To further map different modalities into a unified feature space, Xie et al. [23] extra introduced an immediate text space to minimize the heterogeneity between modalities. Gu et al. [19] enforced the outputs of image representation network to fit the robust outputs of video representation network with the Mean Square Error. Shim et al. [24] integrated image embedding and video embedding into a unified feature embedding with the self-attention mechanism. Porrello et al. [27] devised a teacher-student training strategy to learn more identity-sensitive features against camera view variations. In contrast, our method proposes two key modules to further improve the performance of I2V ReID from the two aspects of solving appearance and modality misalignment. We focus on learning more fine-grained semantic and temporal invariant features for the I2V ReID task. The I2V ReID can also be treated as a special case of video-based person ReID. The video-based person ReID commonly employ the temporal pooling [9,10], optical flow [11], Recurrent Neural Network [12,13] and 3DCNN [28] to mine the temporal motion information of identities. In this work, we also aim to make use of the temporal information in videos, and dedicate to minimize the modality gap existed in I2V ReID.

2.2. Appearance alignment

Existing methods typically address appearance misalignment problem by partitioning human images into several parts to extract more fine-grained local identity features. The partition methods can be roughly divided into the explicit partition and implicit

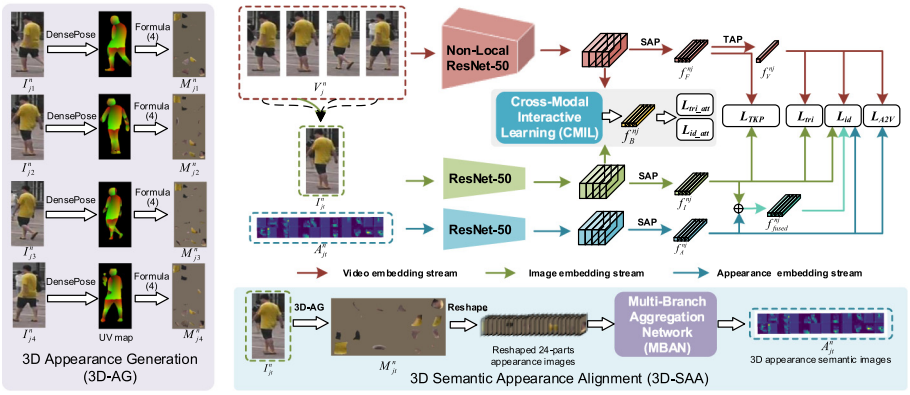

Fig. 2. Overview of the proposed I2V ReID method. There are three main streams in the whole framework, including video, image and appearance embedding streams. These three streams can transform different kinds of inputs into high-level semantic feature space. Specially, a 3D-SAA module is proposed to generate 24-parts local semantic appearance images by the 3D-AG part and MBAN. A CMIL module is also designed to interactively propagate knowledge between two modalities in the feature space. The learned features $f_V^{nj}, f_F^{nj}, f_I^{nj}, f_A^{nj}, f_B^{nj}, f_{fused}^{nj}$ are jointly trained by L_{TKP}, identification-based loss L_{id}, L_{id_att} and verification-based loss $L_{tri}, L_{tri\_att}, L_{A2V}$. (**Best viewed in color**)

partition. The explicit partition leverages extern cues, such as 2D human pose estimation [20], uniform partition [5], and dense human surface [29]. Yu et al. [20] proposed a cross-media body-part attention network for I2V ReID by extracting the cross-modal body part attention features based on the 2D poses of person. Sun et al. [5] proposed a strong baseline which partitions the raw global human images into several uniform stripes to explicitly align body parts of persons. Compared with [20] and [5], our work can capture more detailed local appearance information of persons based on the human 3D parametric model. Zhang et al. [29] also utilized the 3D surface of human to align human body parts for image-based person ReID task. Compared with [29], the proposed 3D-SAA module adopts different part-level feature aggregation manner, different multi-grained identity feature learning method and the crucial body parts selection component.

The essential core of implicit partition is to highlight foreground human body parts with the attention mechanism [7,8,30]. Yao et al. [8] designed a part loss network to highlight a crucial body region and ignore other body regions. Zhao et al. [30] presented a part-aligned representation learned only from person similarities without the supervision information about the human parts. Although these methods generally perform better than the explicit partition based methods, they still focus on the 2D appearance information rather than more detailed 3D information. Our work not only can capture aligned 3D local information of persons but also can weaken the cluttered backgrounds and negligible body parts with the merits of implicit partition.

2.3. Modality alignment

Cross-modal human target analysis, such as RGB-infrared Person ReID [31–33], RGB-D cross-modal person ReID [34], and NLP-based person search [35,36], has been widely developed recently. The RGB-infrared person ReID task is utilized to realize the all-day person re-identification, and the RGB-D cross-modal person ReID task is proposed to ensure the privacy of pedestrians. These cross-modal tasks make person ReID easier to apply to the real scenarios. Compared with these cross-modal ReID tasks mentioned above, the I2V ReID task utilizes the same sensor to capture images or videos as inputs. The modality differences in I2V ReID are less than other cross-modal person ReID tasks, but the I2V ReID task is more practical and flexible in the wild. Although [19] and [24] provide the solutions to address the modality gap in the I2V ReID task, we propose a more effective interactive communication manner between images and videos. By integrating the relation between two modalities into the channels of image features and jointly learning with two embedding streams, the CMIL module can realize the communication between two modalities.

3. The proposed method

This section introduces the proposed deep I2V ReID pipeline to address the appearance and modality misalignment. First, the proposed whole pipeline is introduced and the relation between proposed different components are illustrated. Then, a 3D-SAA module is designed to align different instances of the same identity by mapping to the unified 3D parametric surface model to address the appearance misalignment problem. Finally, a CMIL module is developed to construct the relation of two modalities with an interactive similarity comparison mechanism, and integrate the temporal information of videos to the image features for the joint learning of two modalities.

3.1. The proposed I2V ReID pipeline

Problem Formulation: The I2V ReID belongs to the retrieval task, which needs to match pedestrian candidates between query and gallery set. In contrast to the same modality-based person ReID tasks, the I2V ReID needs to deal with heterogeneous data, eg. query images and gallery videos. Given a query image I_q and a gallery $V_g = \{V_g^i | i \in [1, 2, ..., L]\}$ with L videos, the goal of I2V ReID is to compare with gallery videos V_g based on the identity information in query image I_q, and return a ranked similarity score list $s_{rank} = \{s_{rank}^i | i \in [1, 2, ..., L]\}$. To achieve this goal, it is crucial to extract accurate fine-grained human appearance features as identity information and address the heterogeneity between images and videos. In this work, we propose a 3D-SAA module and a CMIL module to further improve the generalization ability of the I2V pipeline from these two aspects mentioned above.

Overview of Proposed Deep I2V ReID Pipeline: The proposed deep I2V ReID pipeline is depicted in Fig. 2. To learn discriminative identity representations to better match query and gallery in deep semantic feature space, lots of annotated video clips $V_{train} = \{V_j^n | j \in [1, 2, ..., J], n \in [1, 2, ..., N]\}$ are used to train the proposed deep I2V ReID pipeline. The terms J and N denote the number of video clips and human identities, respectively. Each video clip V_j^n in V_{train} contains T frames with the same identity, which is formulated as:

$$V_j^n = \left\{ I_{j1}^n, ..., I_{jt}^n, ..., I_{jT}^n \right\}, t \in [1, T], \quad (1)$$

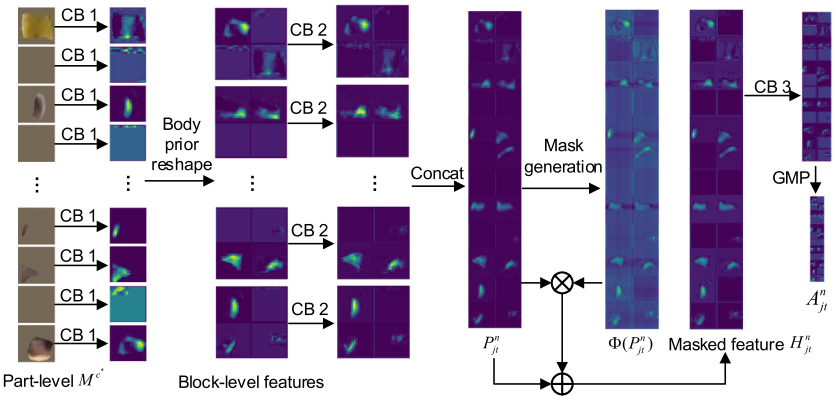

Fig. 3. The architecture of the proposed MBAN. The aligned 24-parts appearance images are transformed into aggregated 3D appearance semantic images with MBAN. The key components in MBAN, like the convolution block (CB) 1, CB 2, CB 3, mask generation part and, body prior reshape part are shown in Fig. 4.

where, $I^n_{jt} \in \mathbb{R}^{3\times W\times H}$ denotes the tth frame in $V^n_j \in \mathbb{R}^{T\times 3\times W\times H}$, and 3, W, H denote the channel number, width and height of I^n_{jt} and V^n_j, respectively.

The proposed I2V ReID pipeline has three main streams, a video embedding stream and an image embedding stream for learning deep features from two modalities, and an additional appearance embedding stream. In the training phase, the video clips are fed into the video embedding stream, and all frames in input video clips are utilized as a image set $\{I^n_{jt}\}$ to train the image embedding stream correspondingly. In the video embedding stream, the ResNet-50 [37] with non-local blocks [38] (Non-Local ResNet-50) and spatial average pooling (SAP) layer is used to extract frame-level features due to its ability to capture temporal information. The extracted frame-level features $f^{nj}_F \in \mathbb{R}^{T\times D\times 1}$ are further integrated to the video-level feature $f^{nj}_V \in \mathbb{R}^{D\times 1}$ by a temporal average pooling (TAP) [19] layer. The symbol D denotes the number of feature channels. In the image embedding stream, each sample in $\{I^n_{jt}\}$ is fed into the ResNet-50 with a SAP layer to learn corresponding image feature $f^{nj}_I \in \mathbb{R}^{T\times D\times 1}$.

To further address appearance misalignment problem, this work proposes a 3D-SAA module to guide the identity feature learning. For each image I^n_{jt}, a 3D appearance generation (3D-AG) part is firstly used to estimate aligned 24-parts appearance image $M^n_{jt} \in \mathbb{R}^{3\times 4s\times 6s}$ of persons. The term s is the width and height of 1-part appearance images. The estimated 24-parts appearance image is reshaped in part-level, and then fed into a designed multi-branch aggregation network (MBAN) to extract aggregated 3D local appearance semantic image $A^n_{jt} \in \mathbb{R}^{3\times 2\tau\times 12\tau}$. The term τ is the width and height of 1-part of A^n_{jt}. The aggregated 3D local appearance semantic image A^n_{jt} belongs to the same identity as I^n_{jt}, but A^n_{jt} contains abundant aligned semantic information than I^n_{jt}. Similar to the image embedding stream, the 3D local appearance semantic image A^n_{jt} is also fed into the ResNet-50 in the appearance embedding stream to extract appearance feature $f^{nj}_A \in \mathbb{R}^{T\times D\times 1}$. The appearance feature f^{nj}_A is further fused with f^{nj}_I to obtained more fine-grained identity features $f^{nj}_{fused} \in \mathbb{R}^{T\times D\times 1}$.

To close the gap between image and video embedding streams, a CMIL module is proposed to model relations between two streams. The outputs of Non-Local ResNet-50 in video embedding stream and ResNet-50 in image embedding stream, are directly fed into CMIL module to interactively propagate knowledge of two modalities, and obtain balanced identity feature $f^{nj}_B \in \mathbb{R}^{T\times D\times 1}$. The feature f^{nj}_B can be used to simultaneously involve the visual information in images and temporal information in videos.

Overall, the learned all features $f^{nj}_V, f^{nj}_F, f^{nj}_I, f^{nj}_A, f^{nj}_B, f^{nj}_{fused}$ mentioned above are jointly trained by multiple loss functions $L_{id}, L_{id\_att}, L_{tri}, L_{tri\_att}, L_{A2V}, L_{TKP}$, as described in Section 3.4.

3.2. 3D semantic appearance alignment

3D Appearance Generation: As shown in Fig. 2, the 3D-SAA module is used to extract aggregated 3D appearance semantic images for guiding the proposed I2V ReID model to learn more semantically aligned identity features. For each image I^n_{jt} in the training phase, it is fed into DensePose model [22] in proposed 3D-AG part to estimate UV map of persons. The UV map [39] can reflect the correspondences between 3D human parametric model [40] and 2D texture maps. Each pixel in I^n_{jt} can be classified into 24 body parts by DensePose model, and transferred to UV coordinates as follows:

$$c^* = argmax_c P(c|I^n_{jt}(x,y)), \tag{2}$$

$$[U,V] = R^{c^*}((x,y)), u = U[x,y], v = V[x,y], \tag{3}$$

where (x,y) is the Cartesian coordinate of the pixel in I^n_{jt}, $P(c|I^n_{jt}(x,y))$ is the probability of pixel $I^n_{jt}(x,y)$ belonging to the cth body part, and c^* is the predicted body part of $I^n_{jt}(x,y)$. A regressor R^{c^*} is used to transform Cartesian coordinates to UV coordinates. Based on the UV map and raw color image $I^n_{jt}(x,y)$, a 24 part-level appearance image M^{c^*} can be extracted by:

$$M^{c^*}[\frac{(\Upsilon - v)s}{\Upsilon}, \frac{(\Upsilon - u)s}{\Upsilon}] = I^{c^*}(x,y), \tag{4}$$

where $I^{c^*}(x,y)$ denotes the region of the c^*th body part in I^n_{jt}, and Υ denotes the maximum value of the UV map. With Formula (4), we can obtain aligned 24-parts appearance image $M^n_{jt} = \{M^{c^*}|c^* \in [1,24]\}$, and 24 body parts are well aligned for further extracting semantic appearance images. The backgrounds of M^n_{jt} are set to the mean value of I^n_{jt} [29], and 24 parts in M^n_{jt} are arranged in the size of 6×4. From Fig. 2, it can be seen that the estimated image M^n_{jt} is uniformly partitioned into body parts in 3D appearance space. This transform can minimize the influence by pose variations and other visual interferences.

Multi-Branch Aggregation Network: Although the extracted appearance image M^n_{jt} by 3D-AG part is aligned along the part-level, the same attention with more crucial body parts, like head and torso, is also paid to some negligible body parts and irrelevant backgrounds. To this end, a MBAN model is designed to deal with different body parts, and learn a mask to calculate the importance of different body parts. The architecture of MBAN is shown in

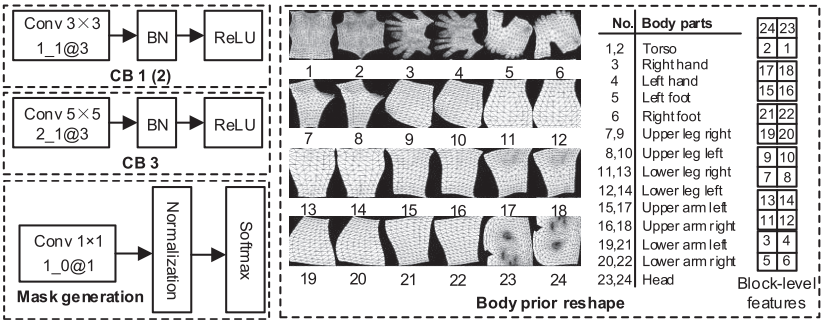

Fig. 4. The key components in MBAN. The CB module contains a convolution layer, a batch normalization (BN) layer and a ReLU layer. Specifically, the "Conv 5×5 2_1@3" in CB 3 denotes the convolution layer with the kernel size of 5×5, the stride of 2, the padding number of 1, and 3 output channels. The meanings of other convolution layers in MBAN are similar to "Conv 5×5 2_1@3". The mask generation part contains a convolution layer, a normalization layer and a Softmax layer. The body prior reshape part shows how the 24 part-level features are concatenated to 6 block-level features.

Fig. 3. Firstly, the aligned 24-parts appearance image M_{jt}^{n} estimated by 3D-AG part is reshaped to the 24 part-level appearance images M^{c^*}. These images are mapped to semantic spaces by convolution block (CB) 1, and reshaped to block-level features with body prior reshape part. As shown in Fig. 4, the body reshape part reshapes the human body parts in the order of head, torso, arms, legs, hands and feet from top to bottom. This kind of reshape operation can minimize the gap between local appearance images and raw global color images with the help of the geometric information of human body structure [41]. Then, after encoding block-level features by CB 2, all block-level features are concatenated to an image-level feature P_{jt}^{n}. The image-level feature P_{jt}^{n} is used to generate a mask $\Phi(P_{jt}^{n})$ by a mask generation part in Fig. 4, and obtain the masked image-level feature H_{jt}^{n} by multiplication and addition operations. The generation process of H_{jt}^{n} is as follows:

$$\Phi[P_{jt}^{n}(x,y)] = \frac{e^{g(P_{jt}^{n}(x,y))/\Psi}}{\sum_{z\in\mathbb{Z}} e^{g(P_{jt}^{n}(z))/\Psi}}, \tag{5}$$

$$H_{jt}^{n} = P_{jt}^{n} + P_{jt}^{n} \star \Phi[P_{jt}^{n}(x,y)], \tag{6}$$

where, $\mathbb{Z} = (x,y)|x = 1,...,2s; y = 1,...,12s, g(\cdot)$ represents a convolution operation of 1×1 kernel size, and $\Psi = ||P_{jt}^{n}||$ is the L2 norm of P_{jt}^{n}. The symbol $\star$ represents the channel-wise Hadamard matrix product operation. The masked image-level features can weaken the influence of negligible body parts and backgrounds. Finally, H_{jt}^{n} is utilized to generate aggregated 3D local appearance semantic image A_{jt}^{n} by CB 3 and a global max pooling (GMP) layer. The reason why we adopt GMP layer rather than global average pooling (GAP) layer here is that the GMP layer is more helpful to only preserve the largest response values for a local view [42].

After the semantic alignment, both I_{jt}^{n} and A_{jt}^{n} will be fed into the weights-irrelevant ResNet-50 model to extract features as f_I^{nj} and f_A^{nj}. The proposed MBAN and ResNet-50 in appearance embedding stream are jointly trained in an end-to-end manner, which can reduce the information loss of MBAN while the identity features are learned. To further guide the learning of f_I^{nj}, we fuse f_I^{nj} and f_A^{nj} by:

$$f_{fused}^{nj} = f_I^{nj} + f_A^{nj}. \tag{7}$$

The fusion between f_I^{nj} and f_A^{nj} as an additional branch is helpful to make global color images and local appearance images compensate each other.

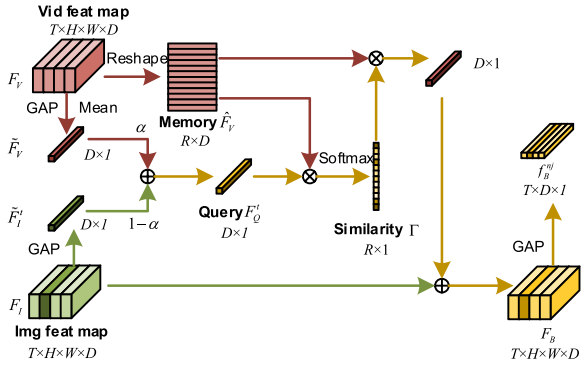

Fig. 5. The architecture of the CMIL module. The "Vid feat map" and "Img feat map" are the frame-level output of Non-local ResNet-50 in video embedding stream and the output of ResNet-50 in image embedding stream, respectively. They are averaged and added to construct a query vector, and the "Vid feat map" is reshaped as a memory matrix. By calculating the similarity between query and memory, a mask can be obtained. Finally a balanced feature is obtained by introducing a residual learning scheme, and pooling by a GAP layer.

3.3. Cross-modal interactive learning

To solve the problem of modality misalignment, this work proposes a CMIL module as shown in Fig. 5. An interactive similarity comparison mechanism in the CMIL module is adopted to build the relation between images and videos by calculating the similarity of features. Compared with the image features, the video features contain the abundant temporal information. To enhance the representation ability of video features, the video features are selected along the spatial and temporal domain by multiplying the similarity. The enhanced video features are integrated into one-dimensional features with the length of channel number, and fused with image features as the balanced identity features. By jointly training the balanced identity features, image features and video features, the gap between two modalities are further minimized.

Specifically, after passing the Non-local ResNet-50 in video embedding stream and ResNet-50 in image embedding stream, we can obtain frame-level feature maps $F_V \in \mathbb{R}^{T\times h\times w\times D}$ of videos, and image feature maps $F_I \in \mathbb{R}^{T\times h\times w\times D}$, respectively. The items h and w are the height and width of feature maps. For each image in $\{I_{jt}^{n}\}$, its feature map is squeezed to an image feature vector $\tilde{F}_I^t$ by a GAP layer. The mean vector $\tilde{F}_V$ of F_V is calculated by a mean operation in temporal domain and a GAP layer in spatial domain. To enhance the relations between image and video modalities, we construct a query vector F_Q^t by the weighted sum operation. The features F_V and F_I are indeed extracted from the same video clip. By fusing

F_V and F_I as the query, the more robust spatial appearance information and temporal motion information are integrated into the query vector for the subsequent effective similarity modeling between two modalities. The query vector F_Q^t is defined as:

$$F_Q^t = \alpha \tilde{F_V} + (1-\alpha)\tilde{F_I^t}, \tag{8}$$

where $\alpha \in [0,1]$ is used to balance the importance of two modalities for constructing query F_Q^t. Moreover, F_V is reshaped to $\mathbb{R}^{R\times D}$ as a memory matrix $\hat{F_V}$, and $R = T \times h \times w$. The construction of query vector and memory matrix is convenient for calculating the similarity between frames in video clips and training images. Based on the $\hat{F_V}$ and F_Q^t, a similarity map Γ is calculated by:

$$\Gamma(\hat{F_V^i}, F_Q^t) = \frac{e^{\hat{F_V^i} F_Q^t}}{\sum_{j\in[1,R]} e^{\hat{F_V^j} F_Q^t}}, \tag{9}$$

where $\Gamma(\hat{F_V^i}, F_Q^t)$ is the ith item in the similarity map. All feature maps in F_I will be repeated by the operations in Formula (8) and Formula (9). Utilizing the learned similarity map Γ and raw image feature map F_I, we can obtain an interactive feature map F_B with a residual learning scheme.

$$F_B = \hat{F_V}^\top \Gamma + F_I, \tag{10}$$

where the term $\hat{F_V}^\top \Gamma$ is repeated as a tensor with the size of $H \times W \times D$ along the spatial domain. The feature map F_B can be further integrated to compact balanced identity feature f_B^{nj} by a GAP layer. The balanced identity feature f_B^{nj} contains the knowledge in both image and video modalities by interactive learning. Then, the feature f_B^{nj} is in conjunction with f_V^{nj} and f_I^{nj} to train the I2V ReID model. The reason why we use $\hat{F_V}$ as the memory rather than image features is that the video modality contains abundant temporal information.

3.4. Joint training

In this paper, the work [19] is utilized as the baseline with the guidance of identification loss L_{id}, triplet loss L_{tri} and temporal knowledge propagation (TKP) loss L_{TKP} [19]. The features f_V^{nj}, and f_I^{nj} are supervised by:

$$L_{id} = -log(\frac{e^{p_{id}^n}}{\sum_{k\in[1,N]} e^{p_{id}^k}}), \tag{11}$$

$$L_{tri} = [m + \max_{f_p \in S_a^+} d(f_a, f_p) - \max_{f_{ng}\in S_a^-} d(f_a, f_{ng})]_+, \tag{12}$$

where $p_{id}^n \in \{w_I^\top f_I^{nj}, w_V^\top f_V^{nj}\}$, and $\{w_I, w_V\}$ represents the classifiers in image and video embedding streams. The term p_{id}^n denotes the classification score of classifying corresponding features into the nth class with the classifier. In Formula (12), the features $\langle f_a, f_p, f_{ng}\rangle$ of identity triplets belong to $\{\langle f_I, f_I, f_I\rangle, \langle f_V, f_V, f_V\rangle, \langle f_I, f_V, f_V\rangle, \langle f_V, f_I, f_I\rangle\}$, m is a pre-defined margin, and $d(\cdot)$ denotes the Euclidean distance. The feature f_a is an anchor in training batch, S_a^+ is positive set with the same identity of f_a, and S_a^- is negative set with the different identity of f_a. In other words, $f_p \in S_a^+$ and $f_{ng} \in S_a^-$ denote the features of positive and negative samples of f_a, respectively. The features f_I and f_V are shorthand for image-level features learned by image embedding stream and video-level features learned by video embedding stream, respectively.

The essence of TKP loss is to minimize the Mean Square Error (MSE) [43] between f_I^{nj} and f_F^{nj}, which is formulated as:

$$L_{TKP} = \frac{1}{NT}\sum_{n,t=1,1}^{N,T} \left\| f_I^{njt} - f_F^{njt} \right\| + \|D_I - D_V\|_F, \tag{13}$$

where $\|\cdot\|$ denotes the square of the L2 norm, and $\|\cdot\|_F$ denotes the square of Frobenius norm. The terms $D_I, D_V \in \mathbb{R}^{NT\times NT}$ represent the Euclidean distance matrices cross different samples in image and video embedding streams.

For f_A^{nj} and f_{fused}^{nj}, the Formula (11) with $p_{id}^n \in \{w_A^\top f_A^{nj}, w_{fused}^\top f_{fused}^{nj}\}$ is also utilized to learn identity-sensitive features, and w_A and w_{fused} are corresponding classifiers. To model the relations between appearance and video embedding streams, a triplet-based loss L_{A2V} is introduced, which is implemented by Formula (12) with $\langle f_A, f_V, f_V\rangle$. The symbol f_A is short for appearance features learned by appearance embedding stream. Moreover, the balanced identity feature f_B^{nj} is supervised by L_{id_att} and L_{tri_att}. The loss L_{id_att} is implemented by L_{id} with $p_{id}^n = w_B^\top f_B^{nj}$, and w_B is the corresponding classifier. The loss L_{tri_att} is implemented by L_{tri} with $\langle f_B, f_B, f_B\rangle$, while f_B is shorthand for balanced features learned by CMIL module. The final loss function is formulated as:

$$L = L_{id} + L_{id\_att} + L_{TKP} + \lambda(L_{tri} + L_{tri\_att} + L_{A2V}), \tag{14}$$

where λ is set to 1.5 following [19]. The larger λ is set for triplet loss than identification loss, since the I2V ReID indeed belongs to verification-based task.

4. Experiments and discussions

4.1. Datasets and settings

MARS Dataset: This dataset [15] contains 1,261 identities and 20,478 tracklets captured by 6 cameras. We follow the baseline [19] to split the dataset into the training and test splits. The training split contains 625 identities and 8,298 tracklets, while the test split contains 635 identities and 11,310 tracklets.

DukeMTMC-VideoReID Dataset: This dataset [10] contains 1,404 identities and 5,534 tracklets captured by 8 cameras in total. We follow the baseline [19] to split the dataset into two splits for training and testing. The training split contains 702 identities and 2,196 tracklets, while the test split contains 702 identities for testing, 408 distractors and 2,636 tracklets.

iLIDS-VID Dataset: This dataset [44] contains 300 identities and 600 tracklets captured by 2 cameras. We follow the baseline [19] to split the dataset into two splits for training and testing. The training split contains 150 identities and 300 tracklets, while the test split contains 150 identities and 300 tracklets.

Implementing Details: In the training phase, we randomly sample 4 frames with a stride of 8 frames from the raw full-length video to form a training video clip. If the length of the raw video is less than 32, we duplicate it to meet the required length. The ResNet-50 and Non-Local ResNet-50 in three different streams are initialized by the ImageNet pre-trained weights and initialization method in [38]. The input video frames are resized to 256×128, and the training batch size is set to 14, which depends on the computation ability of the GPU used in our work. The horizontal flip is used for data augmentation. The parameters $s = 32, \alpha = 0.5$ in Formula (8), and $m = 0.3$ in Formula (12) are set in our work. The dimension D of features is set to 2048. In the backpropagation process, the proposed pipeline is trained with Adam optimizer [45] with weight decay 0.0005, and the initial learning rate is set to 0.0003. In the test phase, we only utilize the raw video and image embedding streams, and the appearance embedding stream is discarded in order not to add extra computation cost and keep the fairness of the evaluation. The setting adopted in the test phase is same as the work [19]. This work is implemented by PyTorch with one NVIDIA GeForce RTX 2080 Ti GPU.

Evaluation Metrics: The task setting [16] is adopted in this work. The Cumulative Matching Characteristics (CMC) [46] and

Table 1
Evaluation of proposed methods with the settings of I2V, I2I and V2V ReID on the MARS dataset. The I2I ReID denotes the image-based person ReID, which is implemented by only using the first frames of query and gallery samples. The V2V ReID denotes the video-based person ReID, which is implemented by using full-length query and gallery videos.

Method	MARS					
	I2V ReID		I2I ReID		V2V ReID	
	top-1	mAP	top-1	mAP	top-1	mAP
TKP (Baseline) [19]	75.6	65.1	71.0	55.0	84.0	73.3
TKP (ReRun)	75.4	64.1	71.1	54.6	82.4	72.2
TKP+3D-SAA	78.5	67.1	72.2	57.0	85.6	75.1
TKP+CMIL	77.3	67.2	71.4	56.1	85.0	75.6
TKP+3D-SAA+CMIL (**Ours**)	**79.1**	**69.0**	**72.6**	**58.1**	**86.1**	**76.9**

Table 2
Evaluation of proposed methods with the settings of I2V, I2I and V2V ReID on the DukeMTMC-VideoReID dataset.

Method	DukeMTMC-VideoReID					
	I2V ReID		I2I ReID		V2V ReID	
	top-1	mAP	top-1	mAP	top-1	mAP
TKP (Baseline) [19]	77.9	75.9	63.4	54.8	94.0	91.7
TKP (ReRun)	76.2	74.2	63.5	54.5	94.0	91.8
TKP+3D-SAA	79.3	77.2	65.5	57.8	93.9	92.2
TKP+CMIL	77.2	76.1	66.7	57.8	94.0	**92.7**
TKP+3D-SAA+CMIL (**Ours**)	**81.2**	**79.1**	**68.4**	**60.0**	**94.9**	92.6

mean Average Precision (mAP) [47] are used to evaluate the performance of proposed methods. These two metrics can well reflect the precision and recall of proposed methods over whole datasets, which has been widely used for many ReID tasks [5,9,48].

4.2. Effectiveness of proposed methods

Settings: The performances of our proposed methods on MARS dataset and DukeMTMC-VideoReID dataset are depicted in Table 1 and Table 2. In our experiments, the work TKP [19] is utilized as our baseline, and all improvements are built on this baseline. Due to the limitation of the computation source, we cannot meet the requirements of training batch size equal to 16 in [19]. The training batch size is set to 14 up to the maximum of the video memory. As shown in Tables 1 and 2, the TKP method is rerun with the batch size of 14, and the results slightly drop compared with the results of "TKP (Baseline)". The "TKP + 3D-SAA", "TKP + CMIL" and "TKP + 3D-SAA + CMIL" denote the baseline with only proposed 3D-SAA module, only proposed CMIL module, and both 3D-SAA and CMIL modules, respectively.

Results of Proposed Methods on I2V ReID: From comparisons between "TKP + 3D-SAA" and baseline under the setting for I2V ReID, it can be seen that the 3D-SAA module has 2.9% top-1 and 2.0% mAP improvements on MARS dataset, and 1.4% top-1 and 1.3% mAP improvements on DukeMTMC-VideoReID dataset. These stable improvements are attributed to the more fine-grained alignment from the 3D-SAA module. From comparisons between "TKP + CMIL" and baseline under the I2V ReID setting, it can be seen that the CMIL module has 1.7% top-1 and 2.1% mAP improvements on MARS dataset. Though "TKP + CMIL" does not have obvious improvements on DukeMTMC-VideoReID dataset, it can work well with the proposed 3D-SAA module. It can be seen that the CMIL module can play an important role due to its interactive communication between two heterogeneous modalities. By comparing "TKP + 3D-SAA + CMIL" with baseline, it can be observed that the fusion of two proposed modules can further improve the performance of I2V ReID on two datasets. These two modules are complementary to each other by addressing two kinds of misalignment problems.

Comparison among I2I, I2V and V2V ReID: Tables 1, 2 and Fig. 6 show the results of proposed methods under different ReID settings, including I2I, I2V and V2V ReID settings. It can be seen that both 3D-SAA and CMIL modules can improve performances under different settings. For all methods, the performances for I2I ReID are worse than I2V ReID, and the performances for I2V ReID are worse than V2V ReID. The reason is that V2V ReID adopts extra temporal information than I2V ReID, and I2V ReID considers abundant visual information than I2I ReID. It is important to address the information loss caused by modality misalignment. As depicted in Fig. 6, our proposed method performs better than baseline under three different ReID settings, since the learned features are more identity-sensitive.

It can be also seen that both 3D-SAA and CMIL modules have certain performance improvements on both mAP and top-1 accuracy. Compared with the method "TKP+3D-SAA", the method "TKP+3D-SAA+CMIL" increases the mAP accuracy by 1.9%, 1.1%, 1.8% under the I2V, I2I and V2V ReID settings, respectively. However, the CMIL module only increases the top-1 accuracy by 0.6%, 0.4%, 0.5% under the I2V, I2I and V2V ReID settings, respectively. By contrast, the CMIL module increases the top-1 accuracy by 0.9%, 2.9%, 1.9% under the I2V, I2I and V2V ReID settings on DukeMTMC-VideoReID dataset, respectively. The differences on two datasets are mainly caused by the dataset characteristics. In [19], the authors give the dataset statistics of two datasets, and the average lengths of person videos in MARS and DukeMTMC-VideoReID datasets are 58 and 168, respectively. The longer person videos help to obtain more robust identity features from person videos. Moreover, the higher mAP performance gain than the top-1 means that the top-1 results are not easily be found on the MARS dataset. Relatively speaking, the shorter person videos in the MARS dataset has more noises than the longer person videos. The proposed CMIL module aims to propagate the temporal motion information from videos to images in order to minimize the modality gap. The temporal motion information in shorter person videos is less than the longer person videos.

4.3. Comparisons with state-of-the-arts

Settings: Tables 3 and 4 show the comparisons between proposed method and state-of-the-art methods on MARS and DukeMTMC-VideoReID datasets, respectively. Table 5 shows the comparisons with other I2V ReID methods on the iLIDS-VID dataset. The comparisons in Tables 3, 4 and 5 are based on the same setting following [19], while the methods with different dataset settings are not reported for fairness. The term "#avgID" mentioned in [24] represents the mean number of identities used in each training batch. As illustrated in [24], the term "#avgID" will influence the performance of the proposed algorithm, so we report the results which have the most similar settings to ours for fairness. To evaluate the effectiveness of the proposed method, we report the results which have the same "#avgID" as the baseline.

Analysis of Competitiveness: From the comparisons in Tables 3 and 4, the proposed method outperforms "P2SNet [16]", "ResNet-50 + XQDA [49]", and "TKP [19]" methods in a large margin on two datasets. The improvements are attributed to the appearance and modality alignment addressed by our proposed I2V ReID pipeline. Compared with the method "READ [24]", our work also performs better than its with the similar settings. The larger "#avgID" is helpful to obtain better results, because more samples with more identities can be introduced to each training batch [24]. However, the results of our proposed method are still competitive with a smaller "#avgID". Although both "TKP" and "READ" contain a cross-modal learning module, they either only transfer knowledge in one direction or integrate two modalities into one intermediate modality. The CMIL module in our pipeline not only

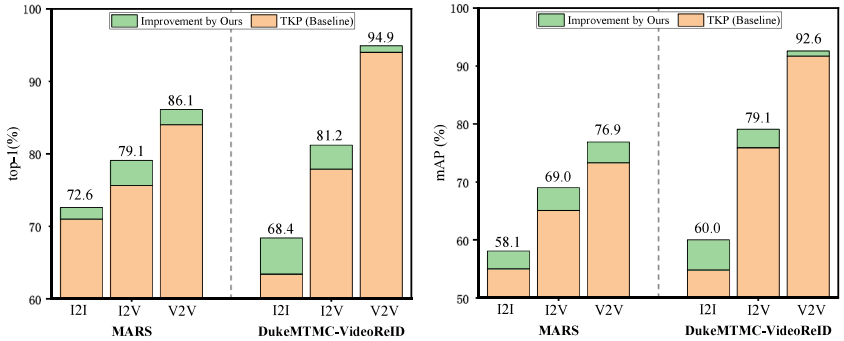

Fig. 6. Comparisons among three settings on MARS and DukeMTMC-VideoReID datasets. The results of TKP baseline, our proposed method, and the improvements are shown.

Table 3
Comparisons between our proposed method and the state-of-arts on MARS dataset.

Method	Source	#avgID	top-1	top-5	top-10	mAP
P2SNet [16]	TCSVT 2017	-	55.3	72.9	78.7	-
ResNet-50 + XQDA [49]	-	-	67.2	81.9	86.1	54.9
TKP (Baseline) [19]	ICCV 2019	4	75.6	87.6	90.9	65.1
TKP (ReRun) [19]	ICCV 2019	4	75.4	87.4	90.7	64.1
DSA [29]	-	4	78.3	88.9	91.4	68.7
STE-NVAN [50]	BMVC 2019	-	80.3	-	-	68.8
NVAN [50]	BMVC 2019	-	80.1	-	-	70.2
MGAT [51]	CVPRW 2019	-	81.1	92.2	-	71.8
READ [24]	ECCV 2020	32	81.5	91.2	93.3	69.9
ResNet-50⋆ [27]	ECCV 2020	8	82.2	-	-	73.4
ResVKD-50 [27]	ECCV 2020	8	83.9	93.2	-	77.3
Ours	-	4	79.1	89.4	91.9	69.0
Ours	-	8	80.2	90.7	92.7	70.8
Ours	-	32	81.3	91.7	93.8	72.6

Table 4
Comparisons between our proposed method and state-of-the-art method on DukeMTMC-VideoReID dataset with the same dataset setting following [19].

Method	Source	#avgID	top-1	top-5	top-10	mAP
TKP (Baseline) [19]	ICCV 2019	4	77.9	-	-	75.9
TKP (ReRun) [19]	ICCV 2019	4	76.2	88.6	91.6	74.2
NVAN [50]	BMVC 2019	-	78.4	-	-	76.7
ResNet-50⋆ [27]	ECCV 2020	8	82.3	-	-	80.2
ResVKD-50 [27]	ECCV 2020	8	85.6	93.9	-	83.8
Ours	-	4	81.2	91.3	93.9	79.1
Ours	-	8	80.9	91.9	93.9	79.4
Ours	-	32	82.8	92.0	94.7	81.0

Table 5
Comparisons with other I2V ReID methods on the iLIDS-VID dataset.

Method	Source	top-1	top-5	top-10	top-20
MPHDL [25]	TIFS 2017	32.6	55.8	69.3	83.2
TMSL [17]	TCSVT 2017	39.5	66.9	79.6	86.6
P2SNet [16]	TCSVT 2017	40.0	68.5	78.1	90.0
Xie et al. [23]	PRL 2020	40.1	67.2	79.7	86.7
TKP (Baseline) [19]	ECCV 2020	54.6	**79.4**	86.9	**93.5**
Ours	-	**54.7**	78.0	**87.3**	92.7

can interactively transfer knowledge between two modalities but also can preserve its own domain knowledge in each modality. Here, we also show the results of state-of-the-art methods, "STE-NVAN" [50], "NVAN" [50], and "MGAT" [51], which validates the effectiveness of the proposed method. The work [27] proposed a self-distillation based method, which is totally different from the motivation of our work. From Tables 3 and 4, it can be seen that the commonly used ResNet-50 model can achieve state-of-the-art performances with the help of [27]. The ResNet-50 model used for teacher model in [27] is called "ResNet-50⋆", while the student model with the view knowledge distillation in [27] is called "ResVKD-50". Although the self-distillation learning manner is not used in our work, the proposed method can still achieve the competitive results compared with [27]. The main contributions in this work are dedicated to provide a solution to learn the identity-sensitive features and address two misalignment problems, rather than strive for achieving the highest mAP and top-K accuracy. The DSA method [29] also introduces the 3D human surface model to the ReID task, however, the motivation and implementation are not the same as our 3D-SAA module. Table 3 shows the results of "DSA" on the MARS dataset by replacing the 3D-SAA module in the proposed pipeline. The improvements verify the effectiveness of the proposed 3D-SAA module.

Compared with MARS and DukeMTMC-VideoReID datasets, the iLIDS-VID dataset has fewer identities and tracklets. The results of the TKP method reported in Table 5 are first pre-trained on the large-scale MARS dataset following [19]. From Table 5, it can be seen that the proposed method can overhead many I2V ReID methods, like MPHDL [25], TMSL [17], P2SNet [16] and Xie et al. [23]. From the comparisons between the baseline and our method, it can be observed that though our method does not achieve the large performance gain, the results are still competitive. The results are caused by the limited training data, while the proposed

Table 6
Analysis of key components in MBAN on MARS dataset.

Method	top-1	top-5	top-10	mAP
TKP (Baseline) [19]	75.6	87.6	90.9	65.1
TKP + 3D-SAA w/o MBAN	76.0	88.4	91.1	65.7
TKP + 3D-SAA w/o Body prior	76.9	89.1	**92.1**	66.7
TKP + 3D-SAA w/o Mask	77.0	88.4	91.2	66.7
TKP + 3D-SAA	**78.5**	**89.2**	92.0	**67.1**

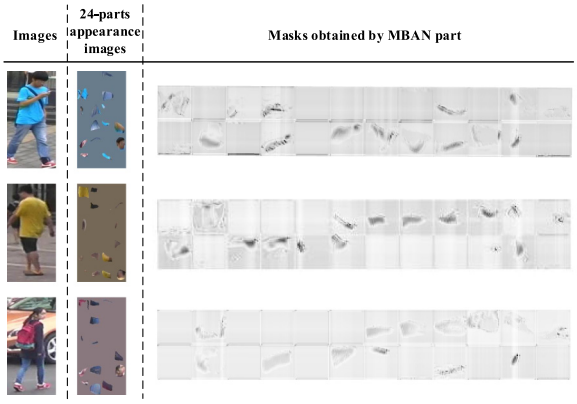

Fig. 7. Some toy examples of raw images, 24-parts appearance images, and masks learned by MBAN. The samples in Row 1 and Row 2 are captured from MARS dataset, while the sample in Row 3 is captured from DukeMTMC-VideoReID dataset.

method has more parameters needing to be trained than the baseline.

4.4. Ablation study

Analysis of MBAN Part: The analysis of key components in MBAN on MARS dataset is depicted in Table 6. From Table 6, it can be seen that the 3D-SAA module improves the mAP and top-1 accuracy by 2.0% and 2.9% in total. If the MBAN is removed, the mAP and top-1 accuracy can only be improved by 0.6% and 0.4%. If the body prior reshape part is removed, the mAP and top-1 accuracy can be improved by 1.6% and 1.3%. The local appearance images with body prior can minimize the gap between appearance branch and image branch with the help of the geometric information of human body structure. Moreover, from the results of "TKP + 3D-SAA w/o Mask", it can be observed that the mAP and top-1 accuracy can only be improved by 1.6% and 1.4%. Fig. 7 shows some examples of learned masks by MBAN. The more dark the color of the mask figure is, the larger the value of the mask is. It can be seen that the learned masks can further decrease the influence of the negligible body parts and backgrounds. By observing the estimated appearance images in Fig. 7, we can also see that the body parts can reflect more detailed appearance information than 2D key joints of the human body. Compared with the improvements obtained from Alp Güler et al. [22], the proposed MBAN and key components in MBAN are more crucial for the performance improvements of 3D-SAA module.

Evaluation of Loss Functions in 3D-SAA Module: The 3D-SAA module is trained with multiple loss functions. Table 7 depicts the influence of different loss functions used for training 3D-SAA module on MARS dataset. The items $L_{id}(f_A)$ and $L_{id}(f_{fused})$ denote the identification loss for features f_A^{nj} and f_{fused}^{nj}, respectively. The item L_{A2V_MSE} means that the loss L_{A2V} is replaced by MSE loss. From Table 7, it can be seen that the fusion of f_I^{nj} and f_A^{nj} results in a performance degradation if the L_{A2V} is not utilized. It implies that the loss L_{A2V} can further guide the fused features to learn more semantic features. If L_{A2V} is replaced by MSE loss, the performance

Table 7
Evaluation of loss functions for 3D-SAA module on MARS dataset.

Loss	Combinations				
$L_{id}(f_A)$	✓	✓	✓	✓	✓
$L_{id}(f_{fused})$		✓		✓	✓
L_{A2V_MSE}				✓	
L_{A2V}			✓		✓
top-1	76.0	75.2	75.9	75.0	**78.5**
mAP	65.3	64.1	66.0	64.5	**67.1**

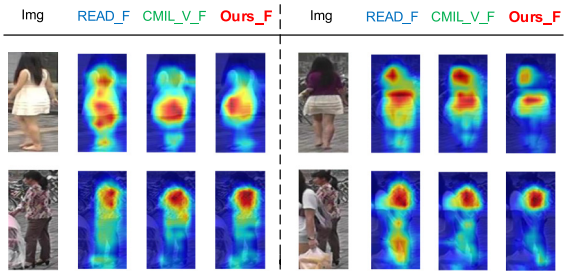

Fig. 8. The visualization of feature maps learned by ResNet-50 in image embedding stream with different communication methods of CMIL module on MARS dataset. The "∗_F" represents the mapping between raw images and heatmaps for the corresponding method "∗". The "READ", "CMIL_V" and "Ours" denote the methods "Ours with READ", "Ours with CMIL_V" and "Ours" in Table 8, respectively. (**Best viewed in color**)

Table 8
Evaluation of different communication ways of CMIL module on MARS dataset.

Method	top-1	top-5	top-10	mAP
Ours with READ	72.9	85.6	89.6	60.5
Ours with CMIL_V	77.0	88.7	91.6	66.7
Ours	**79.1**	**89.4**	**91.9**	**69.0**

with MSE loss cannot achieve the performance of L_{A2V}, since L_{A2V} belongs to a triplet loss which can better constrain distances between videos and appearance images.

Different Communication Ways in CMIL Module: To evaluate the effectiveness of the CMIL module, different communication ways are compared in Table 8. Compared with the method [24], our proposed method interactively communicates between image modality and video modality. However, compared with image modality, the video modality has more abundant temporal information. To this end, an extra branch is added by propagating the learned mutual information to image modality. In Table 8, the method "Ours with READ" means that the CMIL module in our I2V ReID pipeline is replaced with Reciprocal Attention Discriminator (READ) in [24]. We implement this by replacing the fully-connected layer of READ with the GAP layer and adopting the same training batch size as ours, due to the limited computational ability. It can be seen that the READ cannot perform well in our pipeline, when we do not utilize plenties of training identities in each batch, the specially designed sampling strategy, and the reciprocal attention based loss in [24]. The method "Ours with CMIL_V" means the inverse way that it adopts an extra branch by propagating the learned mutual information to video modality. It can be seen that the results of our method are better than "Ours with CMIL_V", since the mutual information can compensate for more information loss of image modality.

Fig. 8 depicts the visualization results of feature maps learned by ResNet-50 in image embedding stream with different communication methods in the CMIL module. It can be observed that the features learned by "Ours with READ" are not centralized enough,

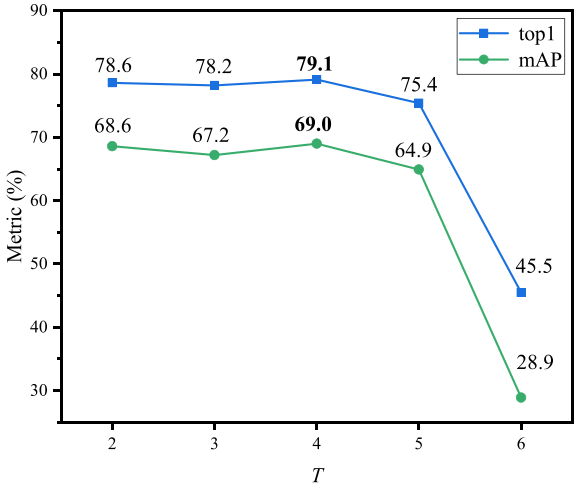

Fig. 9. Performances on variable clip size T conducted with MARS dataset.

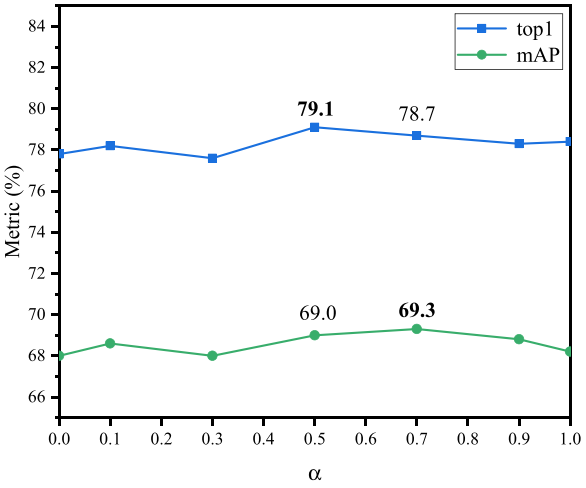

Fig. 10. Evaluation of the parameter α on MARS dataset.

while our method can focus more on the crucial body regions, like the lower torso in Row 1 and upper torso in Row 2. These regions are not easily affected by external interferences, such as clothes changes and occlusion. Compared "Ours with CMIL_V" with ours, it can be seen that the learned features are more fine-grained, since the video clips have more temporal information than images.

Influence by Different Clip Size: Fig. 9 shows the performances on different clip size T conducted with MARS dataset. We evaluate five variants of T from 2 to 6, and the best results are achieved when T is 4. The performance of $T = 6$ drops obviously, since the computation ability of hardware limits the training batch size. The training batch size can only be set to 8 when T is 6, which seriously influences the diversity of training identities in each batch.

Evaluation of the parameter α: The evaluation of the parameter α in Formula (8) is shown in Fig. 10. It can be seen that the top-1 accuracy achieves the best when $\alpha = 0.5$, and the mAP accuracy achieves the best when $\alpha = 0.6$. The results reflect the integration of both image and frame features in videos is more helpful to construct the relation between two modalities. If only image features or frame features are treated as the query in the CMIL module when $\alpha = 0$ or $\alpha = 1$, the query simply preserves the knowledge of one modality, which is not optimal. In this paper, the parameter α is set to 0.5 if not specified.

4.5. Visualization

To further analyze the effectiveness of proposed methods, we also give the visualization results of feature maps learned by ResNet-50 in image embedding stream with different proposed methods, as shown in Fig. 11. Overall, it can be seen that our method focuses on the more robust body-shoulder region of persons. The body-shoulder region is currently considered to be a key cue for representing identity information [52]. By contrast, although the TKP method can learn representative features, the features are still not centralized enough. The 3D-SAA module can focus more on the crucial appearances of persons due to its fine-grained semantic alignment property. The CMIL module is important to find temporal domain invariant information. It can be seen that the features learned by our method are more fine-grained and temporal domain invariant.

In Row 2 and Column 2 in Fig. 11, the person is occluded which results in the appearance misalignment. It can be observed that both the TKP method and our 3D-SAA module focus on the bags of the target-unrelated pedestrian. However, the 3D-SAA module pays more attention on the more robust head-shoulder region. Our CMIL module can ignore the influence of cluttered backgrounds by temporal compensation. From the examples in Row 3, it can also be seen that our proposed method is less susceptible to the changes in the camera view and other appearance information.

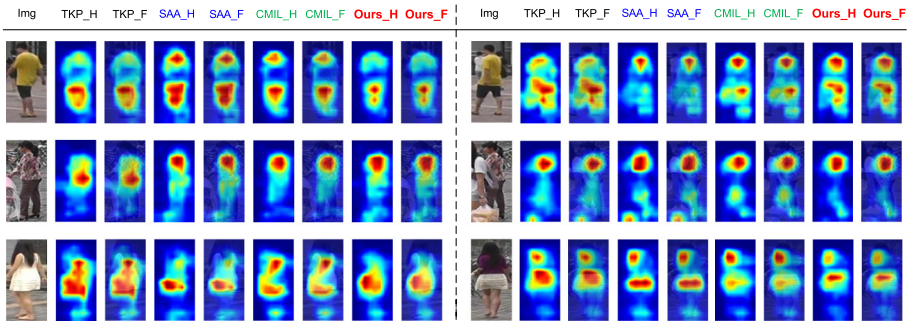

Fig. 11. The visualization of feature maps learned by ResNet-50 in image embedding stream with different proposed methods. The "∗_H" represents the learned heatmap by the corresponding method "∗", while "∗_F" represents the mapping between raw images and heatmaps for the corresponding method "∗". The "TKP", "SAA", "CMIL" and "Ours" denote the baseline TKP method [19], proposed 3D-SAA module, proposed CMIL module, and Our proposed I2V ReID pipeline with both 3D-SAA and CMIL. (**Best viewed in color**)

5. Conclusion

This paper presents a deep I2V ReID pipeline based on proposed 3D-SAA and CMIL modules to address appearance and modality misalignment problems. The 3D-SAA module can semantically align local body parts of persons and weaken the influence of the negligible body parts and cluttered backgrounds. The CMIL module can interactively propagate the modality knowledge of each modality to each other, which can minimize the gap between two modalities. Two complementary modules can guide the deep I2V ReID pipeline to learn more fine-grained and temporal domain invariant feature embedding. This property indicates the generalization ability of our method against misdetections, pose, and camera view variations, for the I2V ReID task. Extensive quantitative and qualitative experiments validate the effectiveness of the proposed method. Although this paper can weaken the influence of the appearance misalignment problem, the proposed I2V ReID pipeline is still affected by the 3D human surface estimation results. The light-weight 3D human surface model, and joint learning of both 3D human surface estimation and I2V ReID can be investigated in the future.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by National Key R&D Program of China (No. 2020AAA0108904), and Science and Technology Plan of Shenzhen (No. JCYJ20190808182209321).

References

[1] N. Jiang, S. Bai, Y. Xu, C. Xing, Z. Zhou, W. Wu, Online inter-camera trajectory association exploiting person re-identification and camera topology, in: Proceedings of the ACM International Conference on Multimedia, 2018, pp. 1457–1465.
[2] Q. Sun, H. Liu, T. Harada, Online growing neural gas for anomaly detection in changing surveillance scenes, Pattern Recognit. 64 (2017) 187–201.
[3] W. Shi, H. Liu, M. Liu, Identity-sensitive loss guided and instance feature boosted deep embedding for person search, Neurocomputing 415 (2020) 1–14.
[4] M. Liu, H. Liu, C. Chen, Enhanced skeleton visualization for view invariant human action recognition, Pattern Recognit. 68 (2017) 346–362.
[5] Y. Sun, L. Zheng, Y. Yang, Q. Tian, S. Wang, Beyond part models: person retrieval with refined part pooling (and a strong convolutional baseline), in: Proceedings of the European Conference on Computer Vision, 2018, pp. 480–496.
[6] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, X. Tang, Spindle net: person re-identification with human body region guided feature decomposition and fusion, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1077–1085.
[7] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, X. Wang, HydraPlus-Net: attentive deep features for pedestrian analysis, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 350–359.
[8] H. Yao, S. Zhang, R. Hong, Y. Zhang, C. Xu, Q. Tian, Deep representation learning with part loss for person re-identification, IEEE Trans. Image Process. 28 (6) (2019) 2860–2871.
[9] S. Li, S. Bak, P. Carr, X. Wang, Diversity regularized spatiotemporal attention for video-based person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 369–378.
[10] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Ouyang, Y. Yang, Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5177–5186.
[11] D. Chung, K. Tahboub, E.J. Delp, A two stream siamese convolutional neural network for person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1983–1991.
[12] Y. Liu, Z. Yuan, W. Zhou, H. Li, Spatial and temporal mutual promotion for video-based person re-identification, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 8786–8793.
[13] N. McLaughlin, J.M. Del Rincon, P. Miller, Recurrent convolutional network for video-based person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1325–1334.
[14] J. Meng, A. Wu, W.-S. Zheng, Deep asymmetric video-based person re-identification, Pattern Recognit. 93 (2019) 430–441.
[15] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, Q. Tian, MARS: a video benchmark for large-scale person re-identification, in: Proceedings of the European Conference on Computer Vision, Springer, 2016, pp. 868–884.
[16] G. Wang, J. Lai, X. Xie, P2SNet: can an image match a video for person re-identification in an end-to-end way? IEEE Trans. Circuits Syst. Video Technol. 28 (10) (2017) 2777–2787.
[17] D. Zhang, W. Wu, H. Cheng, R. Zhang, Z. Dong, Z. Cai, Image-to-video person re-identification with temporally memorized similarity learning, IEEE Trans. Circuits Syst. Video Technol. 28 (10) (2017) 2622–2632.
[18] X. Zhang, S. Li, X.-Y. Jing, F. Ma, C. Zhu, Unsupervised domain adaption for image-to-video person re-identification, Multimed. Tools. Appl. (2020) 1–18.
[19] X. Gu, B. Ma, H. Chang, S. Shan, X. Chen, Temporal knowledge propagation for image-to-video person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 9647–9656.
[20] B. Yu, N. Xu, J. Zhou, Cross-media body-part attention network for image–to-video person re-identification, IEEE Access 7 (2019) 94966–94976.
[21] H. Zhang, J. Cao, G. Lu, W. Ouyang, Z. Sun, DaNet: decompose-and-aggregate network for 3D human shape and pose estimation, in: Proceedings of the ACM International Conference on Multimedia, 2019, pp. 935–944.
[22] R. Alp Güler, N. Neverova, I. Kokkinos, DensePose: dense human pose estimation in the wild, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7297–7306.
[23] Z. Xie, L. Li, X. Zhong, L. Zhong, J. Xiang, Image-to-video person re-identification with cross-modal embeddings, Pattern Recognit. Lett. 133 (2020) 70–76.
[24] M. Shim, H.-I. Ho, J. Kim, D. Wee, Read: reciprocal attention discriminator for image-to-video re-identification, in: Proceedings of the European Conference on Computer Vision, 2020, pp. 335–350.
[25] X. Zhu, X.-Y. Jing, X. You, W. Zuo, S. Shan, W.-S. Zheng, Image to video person re-identification by learning heterogeneous dictionary pair with feature projection matrix, IEEE Trans. Inf. Forensics Secur. 13 (3) (2017) 717–732.
[26] T. Li, L. Sun, C. Han, J. Guo, Salient region-based least-squares log-density gradient clustering for image-to-video person re-identification, IEEE Access 6 (2018) 8638–8648.
[27] A. Porrello, L. Bergamini, S. Calderara, Robust re-identification by multiple views knowledge distillation, in: Proceedings of the European Conference on Computer Vision, Springer, 2020, pp. 93–110.
[28] J. Li, S. Zhang, T. Huang, Multi-scale 3D convolution network for video based person re-identification, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 8618–8625.
[29] Z. Zhang, C. Lan, W. Zeng, Z. Chen, Densely semantically aligned person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 667–676.
[30] L. Zhao, X. Li, Y. Zhuang, J. Wang, Deeply-learned part-aligned representations for person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 3219–3228.
[31] D. Li, X. Wei, X. Hong, Y. Gong, Infrared-visible cross-modal person re-identification with an x modality, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2020, pp. 4610–4617.
[32] M. Ye, X. Lan, J. Li, P.C. Yuen, Hierarchical discriminative learning for visible thermal person re-identification, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2018, pp. 7501–7508.
[33] M. Ye, Z. Wang, X. Lan, P.C. Yuen, Visible thermal person re-identification via dual-constrained top-ranking, in: International Joint Conference on Artificial Intelligence, vol. 1, 2018, pp. 1092–1099.
[34] N. Karianakis, Z. Liu, Y. Chen, S. Soatto, Reinforced temporal attention and split-rate transfer for depth-based person re-identification, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 715–733.
[35] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, X. Wang, Person search with natural language description, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1970–1979.
[36] K. Niu, Y. Huang, L. Wang, Textual dependency embedding for person search by language, in: Proceedings of the ACM International Conference on Multimedia, 2020, pp. 4032–4040.
[37] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[38] X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7794–7803.
[39] T. Alldieck, G. Pons-Moll, C. Theobalt, M. Magnor, Tex2shape: detailed full human body geometry from a single image, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 2293–2303.
[40] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, G. Pons-Moll, Video based reconstruction of 3D people models, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8387–8397.
[41] D. Li, X. Chen, Z. Zhang, K. Huang, Learning deep context-aware features over body and latent parts for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 384–393.
[42] Y. Fu, Y. Wei, Y. Zhou, H. Shi, G. Huang, X. Wang, Z. Yao, T. Huang, Horizontal pyramid matching for person re-identification, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 8295–8302.
[43] K. Han, Y. Huang, C. Song, L. Wang, T. Tan, Adaptive super-resolution for person re-identification with low-resolution images, Pattern Recognit. 114 (2021) 107682.

[44] T. Wang, S. Gong, X. Zhu, S. Wang, Person re-identification by video ranking, in: Proceedings of the European Conference on Computer Vision, Springer, 2014, pp. 688–703.
[45] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, in: International Conference for Learning Representations, 2015.
[46] D. Gray, S. Brennan, H. Tao, Evaluating appearance models for recognition, reacquisition, and tracking, in: Proceeding of IEEE International Workshop on Performance Evaluation for Tracking and Surveillance, vol. 3, Citeseer, 2007, pp. 1–7.
[47] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable person re-identification: a benchmark, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1116–1124.
[48] W. Shi, H. Liu, F. Meng, W. Huang, Instance enhancing loss: deep identity-sensitive feature embedding for person search, in: Proceedings of the IEEE International Conference on Image Processing, IEEE, 2018, pp. 4108–4112.
[49] S. Liao, Y. Hu, X. Zhu, S.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, 2015, pp. 2197–2206.
[50] C.-T. Liu, C.-W. Wu, Y.-C.F. Wang, S.-Y. Chien, Spatially and temporally efficient non-local attention network for video-based person re-identification, in: Proceedings of the British Machine Vision Conference, 2019, pp. 1–13.
[51] L. Bao, B. Ma, H. Chang, X. Chen, Masked graph attention network for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019, pp. 1496–1505.
[52] B. Xu, L. He, X. Liao, W. Liu, Z. Sun, T. Mei, Black Re-ID: a head-shoulder descriptor for the challenging problem of person re-identification, in: Proceedings of the ACM International Conference on Multimedia, 2020, pp. 673–681.

Wei Shi received the B.E. degree in electronic information engineering in 2016. He is working toward the Ph.D. degree in the School of EE&CS, Peking University (PKU), China. His research interests include person re-identification, person search and computer vision. He has already published articles in Pattern Recognition, Neurocomputing, IJCAI, ICASSP and ICIP.

Hong Liu received the Ph.D. degree in mechanical electronics and automation in 1996. He is serving as a Full Professor in the School of EE&CS, Peking University (PKU), China. He has been selected as Chinese Innovation Leading Talent supported by National High-level Talents Special Support Plan since 2013. He is also the Director of Open Lab on Human Robot Interaction, PKU, his research fields include computer vision and robotics, image processing, and pattern recognition. He has published more than 200 papers and gained Chinese National Aerospace Award, Wu Wenjun Award on Artificial Intelligence, Excellence Teaching Award, and Candidates of Top Ten Outstanding Professors in PKU. He is an IEEE member, vice president of Chinese Association for Artificial Intelligent (CAAI), and vice chair of Intelligent Robotics Society of CAAI. He has served as keynote speakers, co-chairs, session chairs, or PC members of many important international conferences, such as IEEE/RSJ IROS, IEEE ROBIO, IEEE SMC and IIHMSP, recently also serves as reviewers for many international journals such as Pattern Recognition, IEEE Transactions on Signal Processing, and IEEE Transactions on Pattern Analysis and Machine Intelligence.

Mengyuan Liu received the Ph.D. degree from the School of EE&CS, Peking University (PKU), China, in 2017, under the supervision of Prof. H. Liu, and served as a Research Fellow of the School of Electrical and Electronic Engineering, Nanyang Technological University (NTU), Singapore, under the supervision of Prof. J. Yuan and Prof. K.K. Ma. He is currently an Associate Professor in Sun Yat-sen University. His research interests include human action recognition and abnormal detection using RGB, depth, and skeleton data. Related methods have been published in T-CSVT, T-MM, PR, CVPR, AAAI, and IJCAI. He has been invited to be a Technical Program Committee (TPC) member for the ACM MM 2018 and 2019. He also serves as a Reviewer for many international journals and conferences, such as the T-II, T-IP, T-CSVT, CVIU, ACM MM, and WACV.