IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 28, 2019, accepted March 17, 2019, date of publication March 27, 2019, date of current version April 11, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2907274

Joint Attentive Spatial-Temporal Feature Aggregation for Video-Based Person Re-Identification

LIN CHEN, HUA YANG, AND ZHIYONG GAO

Department of Electronic Engineering, Institution of Image Communication and Network Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

Corresponding author: Hua Yang (hyang@sjtu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61771303 and Grant 61671289, in part by the Science and Technology Commission of Shanghai Municipality (STCSM) under Grant 17DZ1205602, Grant 18DZ1200102, and Grant 18DZ2270700, and in part by the SJTU-Yitu/Thinkforce Joint Laboratory for Visual Computing and Application, Director Fund of PSRPC.

ABSTRACT Video-based person re-identification (Re-ID) remains to be a promising but challenging computer vision task, suffering from a lack of discriminative features that better aggregate both the spatial and temporal information. In this paper, we propose a joint attentive spatial-temporal feature aggregation network (JAFN) for the video-based person Re-ID, simultaneously learning the quality- and frame-aware model to obtain attention-based spatial-temporal feature aggregation. Specifically, we utilize CNN to learn the spatial features, while introducing the LSTM to separately learn the temporal features. For the feature aggregation, we introduce two attention mechanisms respectively for generating the quality and frame significance score, where the quality score measures the quality of the images for attentive spatial feature aggregation, and the frame score measures the significance of the image frames contributing to the temporal feature. Then, we utilize the set-pooling for both the quality-aware spatial feature and the frame-aware temporal feature aggregation based on the attentive scores. The residual learning is also introduced to play between the LSTM and the CNN for adaptive spatial-temporal feature fusion. Furthermore, we adopt the data balance to alleviate the data disproportions existing in datasets of the video-based Re-ID. The extensive experimental results conducted on the PRID2011, i-LIDS-VID, and MARS datasets demonstrate the effectiveness of the proposed JAFN. Furthermore, comparison results conducted on different modules and features in the JAFN show that our approach is of favorable generalization ability on attentively aggregating both the spatial and temporal features.

INDEX TERMS Person re-identification, attention, spatial-temporal, aggregation.

I. INTRODUCTION

Person re-identification (Re-ID) is of important capability for surveillance systems [1]–[4]. It has remained to be a challenging issue due to multiple challenges, such as large variations in viewpoint and lighting across different views. Most existing approaches have been proposed for the still image case, based on supervised learning to obtain a common spatial feature representation subspace in which the similarity of persons between different views can be assessed directly [5]–[7], or to find distance metric that is helpful for identification in the feature space [8], [9]. However, spatial appearance features are severely limited due to illumination and pose variations among different views. Compared to the methods on still images, video-based Re-ID is a more realistic setting in practice. For the video-based Re-ID task, images in a set can be complementary to each other, so that they provide more information than a single image, such as images from different poses [10], which is useful for reducing the influence of some ambiguous cases and supplying continuous appearance information. Moreover, sequential person images always contain the temporal information such as gait that is more reliable than the appearance information. However, the video-based Re-ID suffers from lack of discriminative features that better aggregate both the spatial and temporal information. Video-based Re-ID remains to be a challenging issue far away from being solved.

The associate editor coordinating the review of this manuscript and approving it for publication was Mohammad Shorif Uddin.

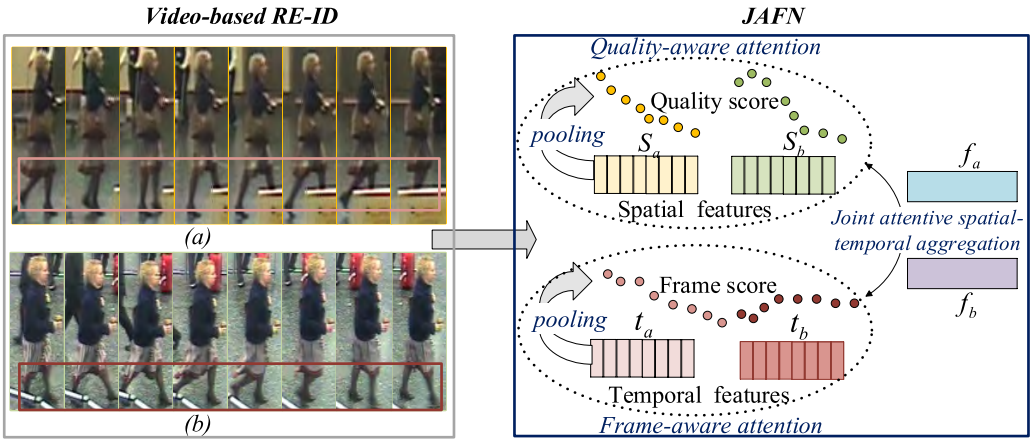


FIGURE 1. The structure of the proposed JAFN to address the challenges for video-based person Re-ID tasks. The JAFN consists of three modules. The quality-aware attention module generates the quality score to measure the quality of the images for attentive spatial feature aggregation. The frame-aware attention module measures the significance of the image frames contributing to the temporal feature. The two attention mechanisms make the system focus on the images with higher quality and significance. The residual learning plays between the spatial and temporal features for more discriminative feature fusion that helps video-based Re-ID.

Recently, many approaches have been proposed for the video-based Re-ID. Some researches propose to direct aggregate information from all person images in a set by the max or average pooling appearance features [11], [12]. However, some images in the set may be not suitable for recognition due to severe clutters and poor definition. To solve the problem, some approaches propose to select the most discriminative frames of the person as the input for recognition [13]. The work of [10] utilizes the quality-aware network model to pay more attention to the images of high-quality. However, these methods only consider the spatial features which easily suffer from the camera view variations. The works of [11] and [12] utilize the deep recurrent network RNN for the video-based person re-identification to extract the temporal feature. The temporal feature is also averagely accumulated from the feature of each frame, ignoring the varied significance of the frames that contribute to the temporal feature learning. The work proposed by [14] utilizes attention model to pay emphasis on the more important area and frames to make the learned feature by the RNN more effective. However, the RNN is known to cannot completely integrate all periodic information of all the sequence frames, the output of which has been proved to easily lose some important information of the earlier person image frames. The temporal feature lacks sufficient appearance information that limits the performance. How to attentively aggregate both the spatial and temporal features remains to be promising and unsolved.

To address the above-mentioned problems, we propose a joint attentive spatial-temporal feature aggregation network (JAFN) for video-based person re-identification. The JAFN is to attentively aggregate both the spatial and temporal feature for obtaining more discriminative feature to improve the performance of video-based Re-ID. As shown in Fig. 1, we propose to learn a quality- and frame-aware model for obtaining attention-based spatial-temporal feature aggregation. Specifically, we utilize the CNN to learn the spatial feature, while introduce the LSTM to separately learn the temporal features. For the feature aggregation, we introduce two attention mechanisms to respectively generate the quality and frame score, while the quality score measures the quality of the images for attentive spatial feature aggregation, and the frame score measures the significance of the image frames contributing to the temporal feature. Then we utilize the set-pooling for both the quality-aware spatial feature and frame-aware temporal feature aggregation based on the attentive scores. For adaptive feature fusion between the two features, we introduce the residual learning played between the LSTM and the CNN for better performance improvement. The element-wise addition is done between the extracted temporal and referenced spatial features for obtaining more discriminative fusion features. We also propose the data balance to alleviate the data disproportions existing in datasets of the video-based Re-ID.

The contributions of our work are summarized as follows: **(1)** Propose a joint attentive feature aggregation mechanism to attentively aggregate both the spatial and temporal features for video-based person re-identification; **(2)** Propose a residual learning mechanism to automatically learn the more discriminative spatial-temporal feature fusion; **(3)** Comprehensive comparisons and discussions are made on different representative datasets to analyze the effectiveness and generalization of our approach.

II. RELATED WORK

Existing approaches for the person Re-ID can be divided into two aspects: feature extraction [15]–[19] or distance learning [8], [9], [20]–[22] to directly learn the projections of data items from different camera views into a common feature representation subspace, in which the similarity between them can be assessed directly [5]–[7], [23].

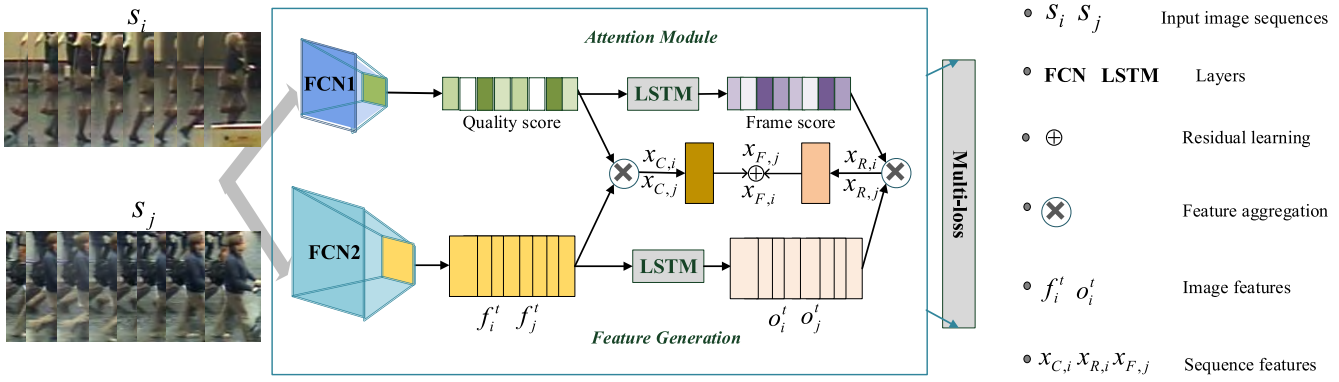


FIGURE 2. The diagram of our proposed JAFN model. It includes attention and feature generation streams for jointly attentive spatial-temporal feature aggregation. We introduce two attention mechanism respectively for generating the quality and frame score, while the quality score measures the quality of the images for attentive spatial feature aggregation, and the frame score measures the significance of the image frames contributing to the temporal feature. Then we utilize the set-pooling for both the quality-aware spatial feature and frame-aware temporal feature aggregation. For adaptive feature fusion between the two features, we introduce the residual learning mechanism played between the LSTM and the CNN for better performance improvement. The element-wise addition is done between the extracted temporal and referenced spatial features for obtaining more discriminative fusion features that help the video-based Re-ID.

However, the over-fitting usually occurs when the system is not of good generality to the out-of-set examples, especially on the challenging datasets which contain large variations across different camera views. To relieve the over-fitting, the work of [10] utilizes the quality-aware network model to pay more attention to the images of high-quality. Zhang et al adopts reinforcement learning to improve the image-level feature representation [24]. Other approaches propose to select the most discriminative frames of the person as the input for recognition [13]. These methods achieve considerably improved performance. However, these methods only consider the spatial features which easily suffer from the camera view variations. In the real surveillance network, large variation always exists in the visual appearance due to clothes variations of the person images cross long period of time or location change, making the performance of identifying persons only by spatial information not reliable enough.

Some representative works introduce the temporal information to help improve the Re-ID performance since it contains more appearance evolution information, such as gait recognition [25]–[28] and temporal sequence matching [1], [29], [30]. Moreover, the works of [11] and [12] utilize the deep recurrent network RNN for the video-based person re-identification to extract the temporal feature. The work proposed by [14] utilizes attention model to pay emphasis on the more important area and frames to make the learned feature by the RNN more effective. These methods achieve considerable improvements of performance since the more reliable temporal feature is utilized. However, the RNN is known to cannot completely integrate all periodic information of all the sequence frames, the output of which has been proved to easily lose some important information of the earlier person image frames. The temporal features extracted from the whole person image sequences easily lose detail of the spatial information. Some researches propose to direct aggregate identity information from all images in a set by simply max/average pooling appearance features of all images [11], [12]. However, some images in the set may be not suitable for recognition due to severe clutters and poor definition. In addition, for the temporal feature, frames in the certain sequence also are of varied significance for the temporal feature extraction. We should pay more attention to the images with better quality and the frames that better help the temporal feature extraction. The spatial and temporal feature should be effectively fused for more performance improvement. He *et al.* [31] proposes a residual learning framework to do element-wise addiction between the deeper layers and the reference to the layer inputs, making the deep layers indirectly better fit a desired optimal mapping by learning the residual. It is easier to learn the residual than to learn the output of the desired layer directly [31]. The performance improvements of the framework on classification demonstrate the effectiveness of residual learning. It is potentially helpful to utilize the residual learning in improving existing deep Re-ID architectures for adaptive spatial-temporal features fusion.

III. METHODOLOGY

A. OVERVIEW

In this section, we concretely introduce the framework of our joint attentive spatial-temporal feature aggregation network (JAFN). To attentively aggregate both the spatial and temporal features for obtaining more discriminative feature for video-based Re-ID, we propose to learn a quality- and frame-aware model to obtain attention-based spatial-temporal feature aggregation. As shown in Fig. 2, The JAFN is of two branches for respective score and feature generation. The score generation branch is to generate the quality and frame score for making the system focus on the feature with more significance, and the feature generation

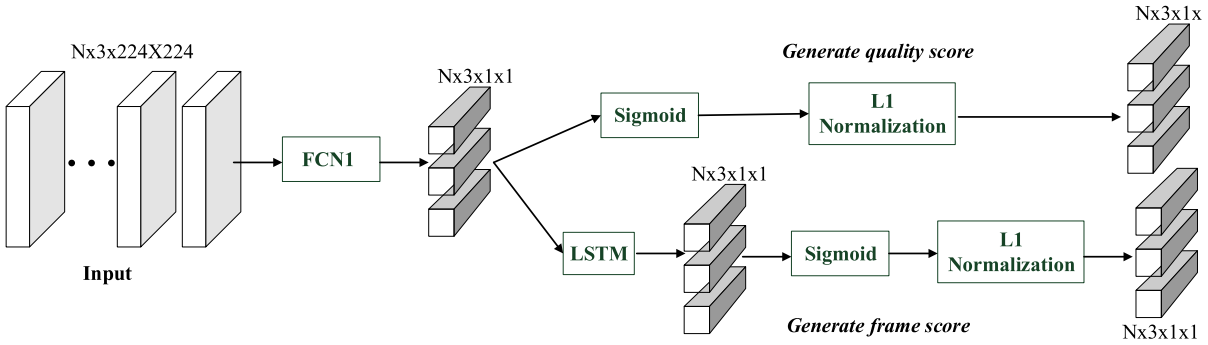


FIGURE 3. The diagram of the quality- and frame-aware score generation. The input data is fed into the convolutional layer "FCN1" to generate 3-dimension score vectors for all the input images. Then the origin quality scores of all images in a set are sent to two parts, one is the sigmoid layer and group L1-normalization layer to generate the final scores for quality measurement, the other through the LSTM to generate the frame score referenced to the output of the quality score for accumulating the significance from anterior images within the video sequence. Furthermore, the 3-dimension score is for the parts attention for more performance improvements.

branch is for respectively generating the spatial and temporal features. Therefore, JAFN mainly contains three parts: the quality-aware attention for spatial feature aggregation, the frame-aware attention for temporal feature aggregation, and the residual learning for spatial-temporal fusion. Moreover, we adopt the data balance to further improve the JAFN.

B. QUALITY-AWARE ATTENTION

The schematic diagram of the quality-aware attention is shown in Fig. 2. The image sequences are passed through two fully convolutional networks (FCN1 and FCN2) to respectively generate the quality score and feature representation. Inspired by the work of [10], the quality-aware attention module aims to measure the usefulness of the input image for the spatial feature aggregation. Intuitively, since images with higher quality are easier for recognition while images with lower quality usually contain less help on set representation, if the image is of high definition and less clutters, the quality score will be theoretically higher to give the feature of the image more attention. The illustration of the quality score generation module is shown in Fig. 3. Specifically, the input data is fed into the convolutional layer "FCN1" to generate 3-dimension score vectors for all the input images. Then the original quality scores of all images in a set are sent to the sigmoid layer and group L1-normalization layer to generate the final scores for quality measurement. The parameters of the "FCN1" is shown on Table 1. Given an input sequence $s = (s^0, \ldots, s^{T-1})$, T is the sequence length. For each person image s^t, the quality score generation can be formulated as:

$$\mu^t = Normalization(Sigmoid((FCN1(s^t)))), \tag{1}$$

where the μ^t is the generated quality-aware score for the image s^t. Meanwhile, the network "FCN2" is for the spatial feature extraction,

$$C(s^t) = FCN2(s^t), \tag{2}$$

where the $FCN2$ denotes the procedure by the fully convolution network shown in Fig. 2, and the $C(s^t)$ denotes the spatial feature after the FCN2. In this paper, the FCN2 of JAFN is a 22 layer googlenet [32]. Denoting the feature after the FCN2 as f^t, then

$$f^t = C(s^t). \tag{3}$$

For the attentive spatial feature aggregation, we adopt the set pooling unit [10] to aggregate the representations of all images with the corresponding quality-aware scores, producing the final quality-aware feature representation of the image set. The procedure can be formulated as:

$$attention(f^t) = \frac{\sum_i^T \mu^t f^t}{\sum_i^T \mu^t}, \tag{4}$$

where the "attention" denotes the set-pooling with generated scores, T is the sequence length for the certain person.

Furthermore, we propose the parts attention mechanism to make the feature focus on the more significant parts. Specifically, we split the global feature into the three-part features, respectively representing the upper part f_1^t, the middle part f_2^t and the lower part f_3^t. The quality-aware score μ^t is 3-dimensional for corresponding parts feature aggregation.

$$X_{C,m} = attention(f_m^t) = \frac{\sum_i^T \mu_m^t f_m^t}{\sum_i^T \mu_m^t}, \tag{5}$$

Then we concatenate the part features to obtain the final feature,

$$X_C = concat(X_{C,m}), \quad m \in 1, 2, 3, \tag{6}$$

TABLE 1. Layer parameters of the "FCN1" shown in Fig. 3 (%).

Name	Type	Number Output	Kernel Size	Stride	Pad
Conv1_s	*convolution*	64	7	2	3
Pool1_s	*Maxpooling*	–	3	2	–
Conv1_ss1	*convolution*	64	3	1	1
Conv2_s	*convolution*	64	3	1	1
Pool_s	*Avepooling*	–	7	7	–
fc1_s	*InnerProduct*	3	–	–	–

where X_C is the final sequential spatial feature after the quality-aware attention module.

C. FRAME-AWARE ATTENTION

The spatial feature usually suffers from the challenges caused by the viewpoint variations. In this section, we propose to attentively aggregate the more reliable temporal features that help the video-based Re-ID. We introduce the recurrent neural network (LSTM) to separately learn the temporal features for the image sequences. Moreover, frames in the certain sequence are also of varied significance for the temporal feature extraction. As shown in Fig. 1, since the temporal feature mainly contains the periodic information such as gait, the images with no clutters on the leg or hand theoretically can provide more stable temporal information, thus these images should be paid more attention. Stimulated by these observations, we propose the frame-aware attention module to obtain the attentive temporal feature. The schematic diagram of the frame-aware attention is also shown in Fig. 2. The LSTM in the JAFN receives the output feature vector from the CNN for accumulating features from anterior images within the video sequence. The input of the LSTM is the feature vectors f^t obtained after the CNN. LSTM learns long-term dependencies and remembers information for long periods of time within person sequence, which can be denoted in the following formula:

$$s^t = sigmoid(W_s[h_{t-1}, f^t] + b_s), \tag{7}$$
$$i^t = sigmoid(W_i[h_{t-1}, f^t] + b_i), \tag{8}$$
$$\widetilde{c}^t = tanh(W_c[h_{t-1}, f^t] + b_c), \tag{9}$$
$$c^t = s^t * c^{t-1} + i^t * \widetilde{c}^t, \tag{10}$$
$$o^t = sigmoid(W_o[h_{t-1}, f^t] + b_o), \tag{11}$$
$$h^t = o^t * tanh(c^t), \tag{12}$$

where the o^t produces an output based on both the current input and information from the previous time-steps, h^t and c^t are the hidden and cell state at time t, f^t is the hidden state of the previous layer at time t or the input for the first layer, the i^t, s^t, and $\widetilde{c}^t$ are the input, forget, and cell gate respectively. Similarly, we utilize the LSTM to generate the frame score v^t referenced to the output of the quality score μ^t for accumulating the significance from anterior images within the video sequence:

$$v^t = sigmoid(W[h_{t-1}, \mu^t] + b), \tag{13}$$

where the v^t is the generated frame-aware score for the frame s^t. We also adopt the set pooling layer for frame-aware sequence feature pooling:

$$attention(o^t) = \frac{\sum_i^T v^t o^t}{\sum_i^T v_t}, \tag{14}$$

Similarly, we adopt the parts attention on the frame-aware attention module to make the learned temporal feature more discriminative. Specifically, we split the temporal feature into the three-part features, respectively representing the upper part o_1^t, the middle part o_2^t and the lower part o_3^t. The frame-aware score v^t is 3-dimensional for corresponding parts feature aggregation.

$$X_{R,m} = attention(o_m^t) = \frac{\sum_i^T v_m^t o_m^t}{\sum_i^T v_m^t}, \tag{15}$$

Then we concatenate the parts features to obtain the final feature,

$$X_R = concat(X_{R,m}), \quad m \in 1, 2, 3 \tag{16}$$

where X_R is the final sequential temporal feature after the frame-aware attention module.

D. RESIDUAL LEARNING MECHANISM

For the video-based Re-ID task, the temporal and spatial appearance features are inextricably intertwined, playing different role to represent the person sequences. To adaptively fuse both the spatial and temporal features, we propose to do cascade residual learning mechanism for improving the video-based Re-ID, where the residual learning is played between the LSTM and the reference of the layer inputs – the CNN, instead of only utilizing CNN or recurrent network to extract features of the pedestrian. As have clarified in the work of [31], the residual learning is to address the degradation problem to make the deep layers indirectly better fit a desired optimal mapping. Formally, denoting the output of the CNN as X, the desired optimal mapping as $H(X)$, in the original network which only contains the CNN network, the original mapping is $X = H(X)$. After the residual learning mechanism, we let the recurrent layers $F(X)$ fit another mapping of $F(X) := H(X) - X$. Then the original mapping is recast into $F(X) + X$. As have clarified in the work [31], it is easier to optimize the residual mapping than to optimize the original unreferenced mapping. To the extreme, if the output of the CNN were optimal, it would be easier to push the residual to zero than to fit an identity mapping by a stack of nonlinear layers. Therefore, the residual learning played by the recurrent layer has the ability to help the CNN to be more optimal and can make the learned feature more discriminative.

Specifically, the JAFN composes of the CNN and the LSTM in a unified network, where the CNN learns the spatial features which are sent to the recurrent neural network (LSTM) for the temporal features learning, and the residual learning is played by the element-wise addition between the spatial and temporal features for obtaining more discriminative fusion features that help the video-based Re-ID. After the residual learning mechanism, we obtain the fusion of the spatial and temporal features x_F:

$$x_F = x_C + x_R, \tag{17}$$

where the "+" denotes the element-wise addition. Then we send the final feature representation for further optimization.

FIGURE 4. The examples of the utilized datasets for the video-based Re-ID problem. Compared to the PRID2011, the i-LIDS-VID is more challenging due to viewpoint variations, cluttered background and occlusions. The MARS dataset is much larger, containing more camera views and tracklets.

E. MULTI-LOSS LAYER

For the JAFN, we adopt the siamese, triplet loss as well as the softmax loss layer for optimization, to fully utilize the label information, pulling positive pairs together and pushing negative pairs apart as well. For the siamese and triplet loss, their both core concept is to divide the input images into pairs, telling the network the image pair is positive or negative. In our case, the positive pair contains three sequences named ''anchor'', ''positive'' and ''negative'', where the ''anchor'' and ''positive'' come from the same person under different camera, the ''negative'' contains the image sequence that comes from the different person under a random camera. Specifically, given an input sequences pair (s_i, s_j), where s_i is the anchor sequence, s_j includes the positive and negative sequences. The feature representation (x_i, x_j) is obtained after our JAFN, then the loss function is calculated in the following formula:

$$L_{sia}(x_i, x_j) = \begin{cases} \frac{1}{2}\left\|x_i - x_j\right\|_2^2, & i = j \\ \frac{1}{2}\left[max(\alpha - \left\|x_i - x_j\right\|_2^2, 0)\right]^2, & i \neq j \end{cases} \tag{18}$$

$$L_{trip}(x_i^a, x_j^p, x_j^n) = \sum_i^N \left[\left\|x_i^a - x_j^p\right\|_2^2 - \left\|x_i^a - x_j^n\right\|_2^2 + \alpha\right] \tag{19}$$

$$L_{sof} = -Wx_{y_i} + log\sum_j e^{Wx_j}, \tag{20}$$

$$L = W_0 L_{trip}(x_i^a, x_j^p, x_j^n) + W_1 L_{sia}(x_i^a, x_j^p, x_j^n) + W_2 L_{sof}(x_i^a) + W_3 L_{sof}(x_j^p) + W_4 L_{sof}(x_j^n), \tag{21}$$

where $\|.\|_2^2$ is the Euclidean distance between the feature vectors. L_{sia} denotes siamese loss while L_{trip} denotes triplet loss, and L_{sof} denotes softmax loss. The final loss is calculated in formula 21. In our case, the α of the siamese loss is 2.0 while the α of the triplet loss is 1.0 to balance the decline rate of the loss and the final accuracy performance. We set the loss weight W_0, W_1,W_2, W_3 and W_4 as 1.0 to ensure correct classification, maximize the relative distance between feature expression of negative pairs, and minimize which between positive pairs.

F. DATA BALANCE

To further improve the performance of the JAFN model, we propose to do data balance to alleviate the data disproportions existing among identities.

In person Re-ID task, there always are hundreds of person images for certain classes while only several images for others in existing datasets. This poses a difficulty in learning algorithms, as they will be biased towards the majority group. To alleviate this disproportions, we propose to balance its identity distribution based on the original images. We enlarge the original datasets based on themselves to make the data distribution balanced, i.e. every identity contains equal person images. Specifically, for a dataset D which contains N identities, where person i contains p_i images, we find the max number p to set the target expanded number, then make up the insufficient sequence of certain pedestrians by copying the original images.

IV. EXPERIMENTS AND RESULTS

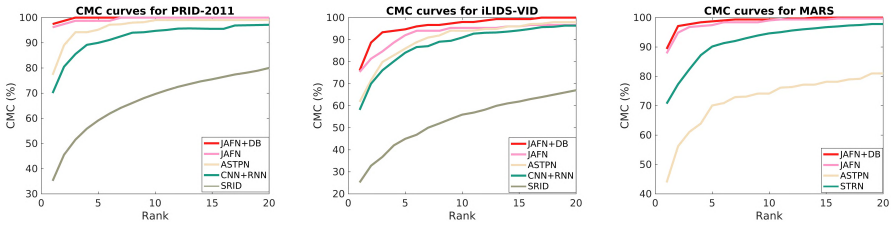
A. EVALUATION DATASETS

In this paper, we adopt three typical datasets widely used on the problem of video-based person re-identification to evaluate the performance of our method. The examples of the utilized datasets are shown in Fig. 4.

PRID-2011: The PRID 2011 [33] re-identification dataset contains two camera views, but only 200 people from the two camera views are adjacent. There is 400 image sequence pair on the dataset. Each image sequence has a variable length from 5 to 675 and with an average number of 100. Compared with the iLIDS-VID dataset, it is captured in uncrowded outdoor scenes, containing more simple and clean background and rare cluttered occlusions.

i-LIDS-VID: The iLIDS-VID dataset [34] is captured at an airport arrival on two non-overlapping camera views. Each camera view contains coincident 300 people and totally 600 image sequence pairs. Each image sequence has a variable length from 23 to 192, and an average number of 73. Due to clothing similarities among people, lighting and viewpoint variations across camera views, cluttered background and occlusions, this dataset is much more challenging than the PRID-2011 dataset.

MARS: The MARS dataset [35] is the largest video re-identification dataset. As an extension of the Market-1501 dataset [36], it consists of 1261 different identities and



(a) (b) (c)

FIGURE 5. Comparison of our method with other state-of-the-art approaches. (a) refers to the CMC rank curves on the PRID-2011 dataset, (b) refers to which on the i-LIDS-VID dataset, (c) refers to which on the MARS dataset.

20715 tracklets. A large number of tracklets contain 25 to 50 frames, while most IDs are captured by 2 to 4 cameras and have 5 to 20 tracklets. Each identity has 13.2 tracklets on average. This dataset is much larger than the previous two datasets, remaining challenging and necessary to evaluate the performance of the proposed method.

B. EXPERIMENTAL SETUP

Input: The input of the improved network after our joint attentive spatial-temporal fusion mechanism is same with the baseline network [10]. We adopt multiple attention mechanisms to excavate the quality- and frame-aware information in the raw images to make the network pay more attention to the more useful images frames.

Network architecture: We first split the middle global feature representation ($1024\times7\times7$) into the different dimensions ($1024\times3\times7$, $1024\times2\times7$, $1024\times2\times7$), then utilize the average pooling for obtaining the 1024-dimension part features. Multiple mechanisms are applied to the part features, including the spatial-, temporal- and parts attention, as well as the residual learning. Therefore, the final feature representation is the concatenation of these attentive parts features, with the dimension of 3072. The learning rate of the JAFN is 1e-3, using the stochastic gradient descent for training.

Data preparation: We randomly select half of the persons for training and the others for testing. For examples, since there are 200 identities in the PRID2011, we randomly select 100 identities for training and the other 100 identities for testing. For the i-LIDS-VID, there are 300 identities, we randomly select 150 identities for training and other 150 identities for testing. Both for training and testing, images are resized to 224×224 for sending to the JAFN.

Training and testing: Each training procedure is 300 epochs. We set the dimension of the feature vector as 3072. During testing, we choose all frames of a certain person and divide it into 8 image frames per-group to extract fusion features with the pre-trained model, and extract final averaged feature to compute the matching rate against the features of gallery sequences under the simple Euclidean distance.

Evaluation metric: We utilize the Cumulative Match Characteristic (CMC) for all the datasets, while extra Mean Average Precision (mAP) [36] for the MARS dataset. The CMC represents the probability that a query identity appears in different-sized candidate lists [3], while the mAP is used together with CMC for the datasets where multiple ground truths from multiple cameras exist for each query.

C. EVALUATION AND RESULTS

1) COMPARISON WITH THE STATE-OF-THE-ART

We compare the performance of the JAFN with several other state-of-the-art video-based methods, shown on Table 2, and the Cumulative Match Characteristic (CMC) curves are also shown in Fig. 5.

From the results, compared to the baseline network [10], JAFN achieves comparable performance improvements. Compared to the baseline proposed by [10] which only utilizes the spatial feature extracted by the CNN, the frame-aware temporal information extracted by the additionally introduced LSTM improves the performance by more than 6% (CMC rank 1) in PRID2011 and 7% (CMC rank 1) in i-LIDS-VID. The performance improvements show the effectiveness of our proposed method in achieving the complementarity between the spatial and temporal features to help the video-based Re-ID. The performance improvement of i-LIDS-VID is more than the PRID2011. We think there are two reasons. One is that the i-LIDS-VID is more challenging than the PRID-2011. Original performance of the PRID2011 is enough considerable (90.3% of CMC rank 1), while the recognition performance for the more challenging i-LIDS-VID is only 68.0% of CMC rank 1, leaving more improvement space for the i-LIDS-VID. The other reason is that the temporal information is more crucial for the i-LIDS-VID than the PRID2011. As can be shown in Fig. 4, the i-LIDS-VID contains much more clutters and variations across different camera views than the PRID2011. Since the spatial feature is limited due to the challenges from cross-view illumination variation, the spatial feature (color, appearance et.al.) is less reliable than the temporal feature (gait et.al.) for the i-LIDS-VID. The original baseline network [10] only contains the CNN to extract the spatial feature, ignoring the significance of the temporal feature. Our JFAN introduces the frame-aware temporal feature to effectively complements the spatial feature, which is more

TABLE 2. Comparison with state-of-the-art methods on different datasets (%).

datasets	*PRID-2011*				*i-LIDS-VID*				*MARS*				
CMC Rank R	R=1	R=5	R=10	R=20	R=1	R=5	R=10	R=20	R=1	R=5	R=10	R=20	mAP
SRID [37]	35	59	70	80	25	45	56	66	–	–	–	–	–
K.Liu et al [38]	64	87	90	92	44	72	84	92	–	–	–	–	–
TDL [39]	57	80	88	94	46	77	90	96	–	–	–	–	–
CNN+RNN [11]	70	90	95	97	58	84	91	96	–	–	–	–	–
ASTPN [14]	77	95	99	99	62	86	94	98	44	70	74	81	–
CRF [23]	77	93	95	98	61	85	94	97	71	89	93	96	–
TSSN [40]	78	94	97	99	60	86	93	97	–	–	–	–	–
RCN [12]	69.0	88.4	93.2	96.4	46.1	76.8	95.6	96.0	–	–	–	–	–
Zheng et al [35]	77.3	93.5	–	99.3	53.0	81.4	–	95.1	68.3	82.6	–	89.4	49.3
STRN [41]	79.4	94.4	–	99.3	55.2	86.5	–	97.0	70.6	90.0	–	97.6	50.7
Zhang et al [24]	85.2	97.1	98.9	99.6	60.2	84.7	91.7	95.2	71.2	85.7	91.8	94.3	–
Baseline [10]	90.3	98.2	99.3	100.0	68.0	86.8	95.5	97.4	–	–	–	–	–
JAFN	96.1	98.7	100.0	100.0	75.3	92.0	95.3	96.7	87.8	97.4	99.0	100.0	73.6
JAFN+DB	**97.4**	**100.0**	**100.0**	**100.0**	**76.0**	**94.7**	**98.0**	**100.0**	**89.3**	**98.7**	**99.3**	**100.0**	**74.9**

crucial for the i-LIDS-VID to achieve favorable performance improvements. The results show that the joint attentive spatial-temporal feature aggregation mechanism can be potentially helpful for being applied to other video-based networks, which do not originally contain the recurrent network for temporal feature extraction. Compared to other state-of-the-art results, our improved networks perform favorably than these approaches, showing the effectiveness of the aggregated feature with our method, especially on the PRID2011 and MARS datasets. After adding the proposed data balance method, denoting as ''+ DB'' in the Table 2, our methods obtain performance improvements on all the datasets, showing that the proposed data balance is helpful for improving the video-based Re-ID.

On the more challenging i-LIDS-VID dataset, the performance (75.3% of CMC rank 1) after our joint attentive feature aggregation mechanism seems more limited than that in PRID-2011 (96.1% of CMC rank 1). One reason is that the PRID-2011 dataset is less challenging than the i-LIDS-VID dataset. Obtaining more performance improvements for the i-LIDS-VID dataset is more challenging. The other reason is that the JAFN contains the CNN and LSTM. In the less challenging PRID2011 and MARS datasets, utilizing the feature from CNN or LSTM can achieve a good performance, then our residual learning mechanism in JAFN for spatial-temporal fusion among these two kinds of features shows more improvement. However, the i-LIDS-VID dataset is much more difficult which contains variations, cluttered background, and occlusions. The extracted spatial and temporal features in the baseline network [11] are not useful enough since training a model on the challenging dataset is more easily to encounter over-fitting, making the residual learning system not work well, which brings the limited performance in i-LIDS-VID.

Moreover, from the results on the work by [35], [36], and [41], the performances of their proposals on the i-LIDS-VID dataset are barely satisfactory compared with the work by [11] while better results of which on the PRID-2011. Our method achieves better performance on these datasets no matter how challenging they are, which also demonstrates the effectiveness and robustness of our approach. Other methods such as [12] and [35] not only utilize the DNN model but also added traditional metric learning method (such as KISSME) to obtain good performance. Our method is evaluated on Euclidean distance and obtain better performance, which shows the effectiveness of the features extracted after JAFN.

2) ANALYSIS OF THE DIFFERENT MODULES

In this section, we further compare the performance between the different modules on JAFN to compare their respective effectiveness. We denote the quality-aware module, the frame-aware module, and the residual learning as ''QA'', ''FA'' and ''RL''. The baseline [10] only utilize the ''QA'' to obtain the model that focuses on the images with higher quality without part attention. We propose the ''FA'' to utilize the LSTM and attention mechanism for obtaining frame-aware temporal feature. Residual learning (''RL'') represents the feature fusion between the spatial- and temporal feature extracted by the ''QA'' and ''RL''. The comparison CMC results are shown on Table 3 and Fig. 6.

TABLE 3. Compare the performance of our different modules in JAFN on different datasets (%).

datasets	*PRID-2011*			*i-LIDS-VID*			*MARS*		
	R=1	R=5	R=20	R=1	R=5	R=20	R=1	R=5	R=20
QA	92.3	100.0	100.0	72.7	86.7	95.3	86.5	97.1	99.7
FA	91.0	98.7	98.7	70.0	85.3	98.0	85.6	96.8	99.7
RL	96.1	98.7	100.0	75.3	92.0	96.7	87.8	97.4	99.7

From the comparison of results, the respective performance of the ''QA'' and ''FA'' is not far different, showing that the spatial and temporal feature should be simultaneously considered for helping recognition. On the other hand, the ''QA'' performs slightly better than the ''FA'' on all the datasets. The results show that since the temporal feature is more difficult to be extracted, the spatial feature is more crucial in recognizing the persons for the Re-ID task. Moreover, the performance of the ''QA'' and ''FA'' all perform favorably than the baseline [10], showing the effectiveness of the frame-aware attention and part attention mechanisms based on the image quality and frame significance. The performance of ''RL'' that combining the ''QA'' and ''FA'' is better than the both, demonstrating the effectiveness of our

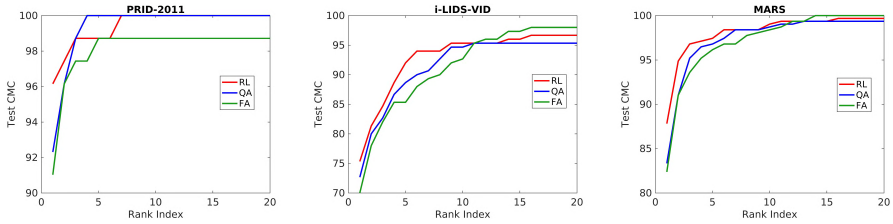


FIGURE 6. Analysis of the effectiveness of the different modules on JAFN. The "QA" refers to the quality-aware module; The "FA" refers to the frame-aware module; The "RL" refers to the residual learning.

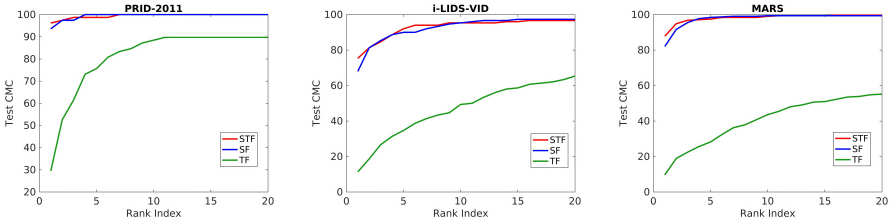


FIGURE 7. Analysis of the effectiveness of the different features in the JAFN. The "SF" refers to the spatial feature extracted by the CNN; The "TF" refers to the temporal feature extracted by the LSTM; The "STF" refers to the fused spatial-temporal feature extracted after the residual learning mechanism.

proposed JAFN utilizing the residual learning mechanism for spatial- and temporal feature fusion.

3) ANALYSIS OF THE DIFFERENT FEATURES

In this section, we further compare the performance of the different features in the JAFN to analyze their respective contributions, that is: the spatial feature extracted by the CNN, the temporal feature extracted by the LSTM, and the fused spatial-temporal feature extracted after the residual learning mechanism. We respectively denote the three kinds of features as "SF", "TF", and "STF". The comparison CMC results are shown on Table 4 and Fig. 7.

TABLE 4. Compare the performance of different features in JAFN on different datasets (%).

datasets	*PRID-2011*			*i-LIDS-VID*			*MARS*		
	R=1	R=5	R=20	R=1	R=5	R=20	R=1	R=5	R=20
SF	93.6	100.0	100.0	68.0	90	97	82.1	98.4	99.4
TF	29.5	75.6	89.7	11.3	34.7	65.3	9.6	28.2	55.1
STF	96.1	98.7	100.0	75.3	92.0	96.7	87.8	97.4	99.7

For the baseline network [10] which only contains the CNN to excavate the quality-aware spatial information, the performance of combining the features extracted by the CNN and LSTM is better than that only from the CNN or LSTM. The results demonstrate the effectiveness of our proposed residual learning mechanism for spatial- and temporal feature fusion, adding the information extracted by the CNN to the LSTM can make up for the lost spatial information from the LSTM. In addition, the performance of the spatial feature extracted by the CNN is much better than that of the LSTM. The CNN dominates in the JAFN. As have clarified in the work of [31], the residual learning framework is to address the degradation problem to make the deep layers indirectly better fit a desired optimal mapping. For the baseline network, after the residual learning mechanism, the introduced recurrent layer helps to learn the residual to make the CNN better fit the desired optimal mapping that helps the recognition. The learned residual with the help of the LSTM is more nearly optimal, which makes the original CNN better close to the desired optimal mapping for better recognition performance. The proposed adaptive residual learning mechanism enhances the significance of the CNN then the brought temporal feature shows limited complementarity, which is demonstrated on the results that the performances of the CNN in the JAFN are better than the baseline [10] on all datasets, as shown on Table 2 and Table 4. Furthermore, the complementarity between spatial features and temporal features brings the crucial help for the video-based Re-ID. The performance of combining the spatial and temporal features is also better than the baseline network [10] only containing the spatial feature. These results shown on the three datasets demonstrate the effectiveness of our proposed JAFN for spatial-temporal feature fusion on the video-based Re-ID task.

V. CONCLUSION

In this paper, we propose a joint attentive spatial-temporal feature aggregation network (JAFN), incorporating quality-aware attention, frame-aware attention and residual learning for improving the video-based person re-identification.

The quality-aware attention module generates the quality score to measure the quality of the images for attentive spatial feature aggregation. The frame-aware attention module measures the significance of the image frames contributing to the temporal feature. The two attention mechanisms make the system focus on the images with higher quality and significance. Then we utilize the residual learning between the spatial and temporal features for more discriminative feature fusion. With comprehensive experiments conducted on the PRID-2011, i-LIDS-VID and MARS datasets, our proposed methods perform favorably against other state-of-the-art networks, demonstrating the effectiveness of JAFN in improving the performance of the video-based Re-ID. We give further comparisons to elaborate the respective advantages and applicability of the different modules and features in JAFN, showing that the proposed parts in JAFN complement each other to help learn more discriminative representations for the video-based Re-ID.

REFERENCES

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

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

[3] L. Zheng, Y. Yang, and A. G. Hauptmann. (2016). "Person re-identification: Past, present and future." [Online]. Available: https://arxiv.org/abs/1610.02984

[4] B. Lavi, M. F. Serj, and I. Ullah. (2018). "Survey on deep learning techniques for person re-identification task." [Online]. Available: https://arxiv.org/abs/1807.05284

[5] H. Zhao *et al.*, "Spindle net: Person re-identification with human body region guided feature decomposition and fusion," in *Proc. CVPR*, Jun. 2017, pp. 907–915.

[6] D. Li, X. Chen, Z. Zhang, and K. Huang, "Learning deep context-aware features over body and latent parts for person re-identification," in *Proc. CVPR*, Jul. 2017, pp. 7398–7407.

[7] L. Zhao, X. Li, J. Wang, and Y. Zhuang, "Deeply-learned part-aligned representations for person re-identification," in *Proc. ICCV*, Oct. 2017, pp. 3239–3248.

[8] J. Lin, L. Ren, J. Lu, J. Feng, and J. Zhou, "Consistent-aware deep learning for person re-identification in a camera network," in *Proc. CVPR*, Jul. 2017, pp. 3396–3405.

[9] H.-X. Yu, A. Wu, and W.-S. Zheng, "Cross-view asymmetric metric learning for unsupervised person re-identification," in *Proc. ICCV*, Oct. 2017, pp. 994–1002.

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

[11] N. McLaughlin, J. M. del Rincon, and P. Miller, "Recurrent convolutional network for video-based person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 1325–1334.

[12] L. Wu, C. Shen, and A. van den Hengel. (2016). "Deep recurrent convolutional networks for video-based person re-identification: An end-to-end approach." [Online]. Available: https://arxiv.org/abs/1606.01609?context=cs

[13] Z. Zhou, Y. Huang, W. Wang, L. Wang, and T. Tan, "See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification," in *Proc. CVPR*, Jul. 2017, pp. 6776–6785.

[14] S. Xu, Y. Cheng, K. Gu, Y. Yang, S. Chang, and P. Zhou, "Jointly attentive spatial-temporal pooling networks for video-based person re-identification," in *Proc. Int. Conf. Comput. Vis.*, Oct. 2017, pp. 4743–4752.

[15] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *Proc. CVPR*, Jun. 2013, pp. 3586–3593.

[16] I. Kviatkovsky, A. Adam, and E. Rivlin, "Color invariants for person reidentification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 7, pp. 1622–1634, Jul. 2013.

[17] R. Zhao, W. Ouyang, and X. Wang, "Person re-identification by salience matching," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2528–2535.

[18] Y. Li, Z. Wu, S. Karanam, and R. J. Radke, "Multi-shot human re-identification using adaptive fisher discriminant analysis," in *Proc. Brit. Mach. Vis. Conf.*, 2015, p. 2.

[19] L. Lin, H. Luo, R. Huang, and M. Ye, "Recurrent models of visual co-attention for person re-identification," *IEEE Access*, vol. 7, pp. 8865–8875, 2019.

[20] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Jun. 2006, pp. 1735–1742.

[21] S. Ding, L. Lin, G. Wang, and H. Chao, "Deep feature learning with relative distance comparison for person re-identification," *Pattern Recognit.*, vol. 48, no. 10, pp. 2993–3003, Oct. 2015.

[22] T. Ni, Z. Ding, F. Chen, and H. Wang, "Relative distance metric leaning based on clustering centralization and projection vectors learning for person re-identification," *IEEE Access*, vol. 6, pp. 11405–11411, 2018.

[23] L. Chen, H. Yang, J. Zhu, Q. Zhou, S. Wu, and Z. Gao, "Deep spatial-temporal fusion network for video-based person re-identification," in *Proc. CVPRW*, Jul. 2017, pp. 1478–1485.

[24] J. Zhang, N. Wang, and L. Zhang, "Multi-shot pedestrian re-identification via sequential decision making," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6781–6789.

[25] J. Man and B. Bhanu, "Individual recognition using gait energy image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 2, pp. 316–322, Feb. 2006.

[26] R. Martín-Fèlez and T. Xiang, "Gait recognition by ranking," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2012, pp. 328–341.

[27] M. S. Nixon, T. Tan, and R. Chellappa, *Human Identification Based on Gait*, vol. 4. Springer, 2010.

[28] S. Sarkar, P. J. Phillips, Z. Liu, I. R. Vega, P. Grother, and K. W. Bowyer, "The humanID gait challenge problem: Data sets, performance, and analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 2, pp. 162–177, Feb. 2005.

[29] Z. Lin, Z. Jiang, and L. S. Davis, "Recognizing actions by shape-motion prototype trees," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 444–451.

[30] D. Simonnet, M. Lewandowski, S. A. Velastin, J. Orwell, and E. Turkbeyler, "Re-identification of pedestrians in crowds using dynamic time warping," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2012, pp. 423–432.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[32] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. CVPR*, Jun. 2015, pp. 1–9.

[33] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, *Person Re-Identification by Descriptive and Discriminative Classification*. Berlin, Germany: Springer, 2011.

[34] W. S. Zheng, S. Gong, and T. Xiang, "Associating groups of people," in *Proc. Active Range Imag. Dataset Indoor Surveill.*, 2009.

[35] L. Zheng *et al.*, "MARS: A video benchmark for large-scale person re-identification," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 868–884.

[36] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1116–1124.

[37] S. Karanam, Y. Li, and R. J. Radke, "Sparse re-id: Block sparsity for person re-identification," in *Proc. CVPRW*, Jun. 2015, pp. 33–40.

[38] K. Liu, B. Ma, W. Zhang, and R. Huang, "A spatio-temporal appearance representation for video-based pedestrian re-identification," in *Proc. ICCV*, Dec. 2015, pp. 3810–3818.

[39] J. You, A. Wu, X. Li, and W.-S. Zheng, "Top-push video-based person re-identification," in *Proc. CVPR*, Jun. 2016, pp. 1345–1353.

[40] D. Chung, K. Tahboub, and E. J. Delp, "A two stream siamese convolutional neural network for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 1992–2000.

[41] Z. Zhou, Y. Huang, W. Wang, L. Wang, and T. Tan, "See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 6776–6785.

LIN CHEN received the B.Sc. degree *(summa cum laude)* in electronic information science and technology from Nankai University (NKU), Tianjin, China, in 2015. She is currently pursuing the Ph.D. degree with the Department of Electronic Engineering, Institution of Image Communication and Network Engineering, Shanghai Jiao Tong University (SJTU), Shanghai, China.

Her research interests include computer vision, deep learning, video surveillance, and person re-identification. She received the 1st Award of the Wider Challenge Person Search on ECCV2018 and the Best Poster Award on IFTC2018.

HUA YANG received the Ph.D. degree in communication and information from Shanghai Jiao Tong University (SJTU), China, in 2004, and the B.S. and M.S. degrees in communication and information from Harbin Engineering University, China, in 1998 and 2001, respectively.

She is currently an Associate Professor with the Department of Electronic Engineering, SJTU. Her current research interests include video coding and networking, computer vision, and smart video surveillance. She was the Supervisor of the 1st Award Team for Wider Challenge Person Search on ECCV2018 and received the Best Poster Award on IFTC2018.

ZHIYONG GAO received the B.S. and M.S. degrees in electrical engineering from the Changsha Institute of Technology (CIT), Changsha, China, in 1981 and 1984, respectively, and the Ph.D. degree from Tsinghua University, Beijing, China, in 1989. From 1994 to 2010, he took several senior technical positions in U.K., including a Principal Engineer with Snell & Wilcox, Petersfield, U.K., from 1995 to 2000, a Video Architect with 3DLabs, Egham, U.K., from 2000 to 2001, a Consultant Engineer with the Sony European Semiconductor Design Center, Basingstoke, U.K., from 2001 to 2004, and a Digital Video Architect with Imagination Technologies, Kings Langley, U.K., from 2004 to 2010. Since 2010, he has been a Professor with Shanghai Jiao Tong University. His research interests include video processing and its implementation, video coding, digital TV, and broadcasting.