Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/mbe

MBE, 20(2): 3504–3527.
DOI: 10.3934/mbe.2023164
Received: 30 September 2022
Revised: 12 November 2022
Accepted: 16 November 2022
Published: 06 December 2022

Research article

Video-based Person re-identification with parallel correction and fusion of pedestrian area features

Liang She[1,2], Meiyue You[3], Jianyuan Wang[4,*] and Yangyan Zeng[5,*]

[1] School of Computer Science and Engineering, Central South University, Changsha 410083, China

[2] School of Computer Science, Hunan University of Technology and Business, Changsha 410205, China

[3] School of Computer Science and Engineering, Beihang University, Beijing 100191, China

[4] School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China

[5] School of Frontier Crossover Studies, Hunan University of Technology and Business, Changsha 410205, China

* **Correspondence:** Email: wangjy90@buaa.edu.cn, yangyanz0930@163.com.

Abstract: Deep learning has provided powerful support for person re-identification (person re-id) over the years, and superior performance has been achieved by state-of-the-art. While under practical application scenarios such as public monitoring, the cameras' resolutions are usually 720p, the captured pedestrian areas tend to be closer to 128×64 small pixel size. Research on person re-id at 128×64 small pixel size is limited by less effective pixel information. The frame image qualities are degraded and inter-frame information complementation requires a more careful selection of beneficial frames. Meanwhile, there are various large differences in person images, such as misalignment and image noise, which are harder to distinguish from person information at the small size, and eliminating a specific sub-variance is still not robust enough. The Person Feature Correction and Fusion Network (FCFNet) proposed in this paper introduces three sub-modules, which strive to extract discriminate video-level features from the perspectives of "using complementary valid information between frames" and "correcting large variances of person features". The inter-frame attention mechanism is introduced through frame quality assessment, guiding informative features to dominate the fusion process and generating a preliminary frame quality score to filter low-quality frames. Two other feature correction modules are fitted to optimize the model's ability to perceive information from small-sized images. The experiments on four benchmark datasets confirm the effectiveness of FCFNet.

Keywords: deep learning; person re-identification; feature fusion; alignment; pixel attention

1. Introduction

Video surveillance is an important part of the security protection system. By aggregating surveillance video data together, a large area with multiple cameras can be monitored. However, with the rapid growth of monitoring data, limited professional identification personnel even have difficulty in analyzing and processing large amounts of data with high precision. The drawbacks of inaccurate and inefficient traditional manual visual inspection are more obvious. Under the modern situation of increasing demand for security, intelligent analysis of video content based on deep learning can effectively improve the efficiency of the monitoring system.

The whole body information of pedestrians, which is collected by cameras at the top view angle, can be used to lock and identify pedestrians. But there is generally no overlap between multiple cameras during the deployment of image acquisition devices for practical application scenarios. The research of Person Re-Identification (person re-id) [1, 2] in the field of computer vision enables cross-camera tracking of a given pedestrian, thus solving the problem that person retrieval becomes more difficult without continuous field-of-view information. The vigorous development of deep learning enables person re-id processing models to demonstrate more accurate decision-making power in an optimized process, which can adapt to the rapid growth of mass monitoring video data [3–5].

Based on surveillance video data, many excellent person re-id research works emerge, they mainly focus on two key points: extracting image spatial information and promoting temporal information cooperation. In order to extract pedestrian features from images adequately, many methods have to pay attention to how to deal with the variances of person re-id, such as handling pedestrian area misalignment [6–8], resolving the noise of image [9–11], adapting to imaging condition differences [12–14], and so on. The methods of mining local details of pedestrians also show their advantages [15–18]. To exert the informative value of video sequences, many methods to study temporal relationships between frames were developed to fuse frame-level features, such as utilizing RNN [19–21], temporal attention [22–24], optical flow [25, 26], and so on. Most outdoor surveillance cameras capture the entire scene at a megapixel resolution ($1280 \times 720 = 921{,}600$). As shown in Figure 1, the pedestrian areas are detected and clipped from the scene with a pixel height of approximately half the common person re-id pixel height of 256, closer to 128×64. Fewer pixels mean less effective information, and the overall quality of frames in the entire video sequence is decreased, it is more difficult to fully distinguish interfering pixels from pedestrian pixels, and not all frames can provide beneficial information for feature construction.

(a) 86×39 **(b)** 103×45 **(c)** 142×55

Figure 1. Examples of video images collected from real-world surveillance. According to the pixel sizes of person areas indicated under each subgraph, the heights of person areas marked by red boxes are closer to 128.

In this paper, we propose a Person Feature Correction and Fusion Network (FCFNet), which can extract person spatial features better by eliminating both misalignment and noise interference as much as possible under the limited information of small-sized images. On the other hand, the overall reduction of information also makes some low-quality frame images unable to provide more help to enhance the robustness of person features. The total amount of these low-quality frame images cannot be underestimated, which wastes valuable storage resources and searching time. We study filtering low-quality frames to improve the proportion of important information while ensuring recognition accuracy. The combination of person feature correction and frame filtering promotes the integration of valuable information in the spatial dimension of frames and the temporal dimension of the entire video. In conclusion, the main contributions of this paper are threefold:

1) For the monitoring scenes with less pixel information, we guide the redundant backgrounds deletion of 128×64 size person images based on affine transformation to reduce the entanglement of unrelated information with subsequent re-id features.
2) We introduce a pixel-level attention mechanism to enhance the attention of irregular pedestrian areas, reduce the interference of noise, and mine more discriminative features in the spatial dimension.
3) We utilize the inter-frame attention mechanism to obtain the importance score of each frame image in the video sequence, which ensures the identification accuracy while reducing resource allocation on frames that have little help in feature representation.

2. Related work

The critical points of video-based re-id are to extract robust person representations in both spatial and temporal dimensions. We succinctly present typical works closely related to ours in this section from two aspects: extracting image spatial information and promoting temporal information association.

2.1. Extracting image spatial information

The video sequence is essentially composed of multiple frame images. Some image-based person re-id works also provide rich ideas for extracting spatial features. Early on, there are many hand-crafted feature representation-based approaches have been proposed to deal with various conditions [27–30], mainly focusing on investigating pose, background, viewpoint, and illumination changes.

The maturity of deep learning is complemented by the emergence of large-scale datasets [3–5], the methods combined with deep neural network shows their ability to extract high-level features. Representation learning and metric learning are the two most basic directions for person re-id [1, 2], their loss measurement ideas are different. The representation learning based methods [31, 32] can automatically extract representation features from original images according to mission objectives through neural networks, and deal with person identification from the perspective of verification or classification. Geng et al. [31] calculated the loss of classification error based on the predicted ID, and judged whether two pictures belong to the same pedestrian to obtain verification loss. And the metric learning based methods [33, 34] often use loss measures such as Contrastive Loss and Triplet Loss to reduce the similarity between different person features and the diversity of the same person features.

On the basis of these two directions, many studies have explored different challenges in mining richer person information.

To learn a sufficiently generalizable model, some studies [15, 17, 35] additionally utilized the attributes of pedestrian pictures, such as gender, hair, clothing, and so on. Lin et al. [15] used output features to simultaneously predict the ID information and various pedestrian attributes of persons, which enhances the generalization ability of the network. Li et al. [17] explored the alignment between human attributes and corresponding local areas, learned the semantic information of the attributes that remain unchanged in multiple domains. Considering the image variances due to noise, misalignments and different perspectives, and so on. Hou et al. [16] proposed BiCnet with two branches, each appended with multiple parallel spatial attention modules, focusing on different complementary regions in consecutive frames. Wang et al. [11] semantically aligned the active regions corresponding to specific human body key points in multiple channels of different images, suppressing the negative impact of occlusion on global feature extraction. Wang et al. [8] proposed a framework containing three branches, inserted a full attention block into an attention branch combining foreground and background information to create attention information in channel direction and spatial direction to adapt to pedestrian misalignment. There are also methods [6, 7] for aligning pedestrian regions based on affine transformations. Zheng et al. [36] used verification loss to constrain the training and proposed an attention-driven two-branch structure model to adapt to spatial positioning and viewpoint variance. Li et al. [14] considered that the features of the same pedestrian under different viewpoints have different discrimination abilities, inferred aggregated multi-view features from single-view images, and achieved a comprehensive description of pedestrian appearance to improve recognition performance. Li et al. [13] further focused on the risks of privacy leakage and data loss that could arise when addressing the diversity of camera views and body postures. To capture the nuances of pedestrian differences, some studies focus on dividing the entire feature map by predefined strips [37, 38], combining local features with global features [39], and so on. Wang et al. [18] fully aggregated the local features from the same body part to obtain corresponding class markers, enabling domain alignment and classification of human body parts without pseudo-label prediction. In addition, Zhu et al. [12] focused on different camera imaging conditions and light variation differences to solve individual cross-domain problems.

With less pixel point information, it is more laborious to distinguish between interfering pixels and pedestrian pixels in a small range of sensing fields. After multilayer convolution, the person information becomes more entangled with unrelated information in the deep features. Eliminating specific variance is not robust enough in harsh scenarios.

2.2. Promoting temporal information association

Compared with several pictures, video sequences hold more spatiotemporal information for person re-id. Video-based person re-id further studies how to better fuse frame-level features based on extracting spatial information of images. Common time-domain modeling methods for aggregating inter-frame features are temporal pooling, RNN [19–21], temporal Attention [22–24], 3D Convolution [40] and so on.

McLaughlin et al. [19] adopted CNN-RNN network architecture, took RGB and optical flow information as inputs, attached RNN to the convolution layer to allow time information to flow in time steps, and then aggregated features of all time steps based on temporal pooling. Rahman et

al. [24] proposed a time domain model based on the fully convolutional network to generate the attention score of video frames. They treated time attention as a sequential annotation problem. The final feature vector of video level is the average of attention pooling and conventional temporal pooling. Zhao et al. [23] re-weighted the frame features based on person attribute groups. Similar to Lin et al. [15], they designed a multi-task network to recognize the pedestrian ID and attributes simultaneously, learned the weights of each frame based on different attribute groups, and concatenated the fused grouping features and global features as re-id metrics. Gao et al. [22] proposed an attention generation network based on time convolution to extract inter-frame information. The extracted spatial feature sequence is taken as the input of the temporal convolutional layer, and the temporal attention of inter-frame features can be obtained by using a concise and effective attention generation structure. Li et al. [41] proposed combining short-term and long-term modeling when fusing the features of frames. Dilated Convolution is used to convolute the adjacent frames of the current frame together to achieve short-term modeling. Based on the relationship between frames, the information contained in all frames is selectively given to the current frame to simulate long-term modeling.

Although simple in structure, temporal convolution can provide effective help for the weighted fusion of frame-level features. Most research methods consider fully integrating the entire video sequence information. However, with the reduction of pixel size and overall degradation of video frame quality, many frame images may not be able to provide help for the formation of person features with high recognition, or even interfere with the feature expression.

3. Person feature correction and fusion network

In actual video surveillance scenes, the captured images tend to have smaller pixel sizes. Compared with the commonly used 256×128 sizes, all frames have less discriminative information, and contain more interferences that are difficult to distinguish from pedestrian pixels, such as misalignment and image noise. In this section, we first demonstrate an overview of our FCFNet, then present the design of each sub-modules from two aspects: correcting the frame-level features and fusing high-quality features.

3.1. Overview of the proposed method

The entire flowchart shown in Figure 2 includes a backbone, two processes, and three sub-modules. The first process is the pre-training based on the backbone ResNet-50 [42] to generate frame quality scores, which can help us filter frames suitable for the overall video-level feature expression. The Frame Quality Assessment Module (FQAM) uses temporal convolution to introduce an inter-frame attention mechanism, which weights the frame-level features "Res5f" of a video in parallel, enhances the comparability of features, and promotes the dominant role of features with rich content information. The convolutional output obtained by this module can reflect the frame quality from specific values after numerical scaling. Then, we ensure the model detection capability by filtering the appropriate proportion of low-quality frames and conducting subsequent experiments using lightweight datasets.

The second process is to insert two sub-modules on the basis of backbone to compel the feature correction of person areas, we use the initial uncorrected features "Res2c" and "Res4f" output from the second stage "Conv2" and fourth stage "Conv4" of the ResNet-50 in the first process to eliminate

the variances, reducing the resource occupation of irrelevant information in the feature construction process. Inspired by the success of Spatial Transformer Networks (STN) [43] in classification tasks, we designed a global alignment module (GAM), which uses high-level semantic features "Res4f" to guide the affine transformation from shallow "Res2c" to align pedestrian areas, and eliminates most of the redundant background interference at the front of the model input. To further adapt to irregular pedestrian areas, we develop the Pixel-Wised Attention Module (PWAM), which introduces the pixel-level attention mechanism to focus the model on the human body. With the cooperation of GAM and PWAM in front and back of the network, features that are helpful for pedestrian identification can be extracted from small-size images and then fused into video-level features by FQAM to give full play to the data advantage of video-based re-id.

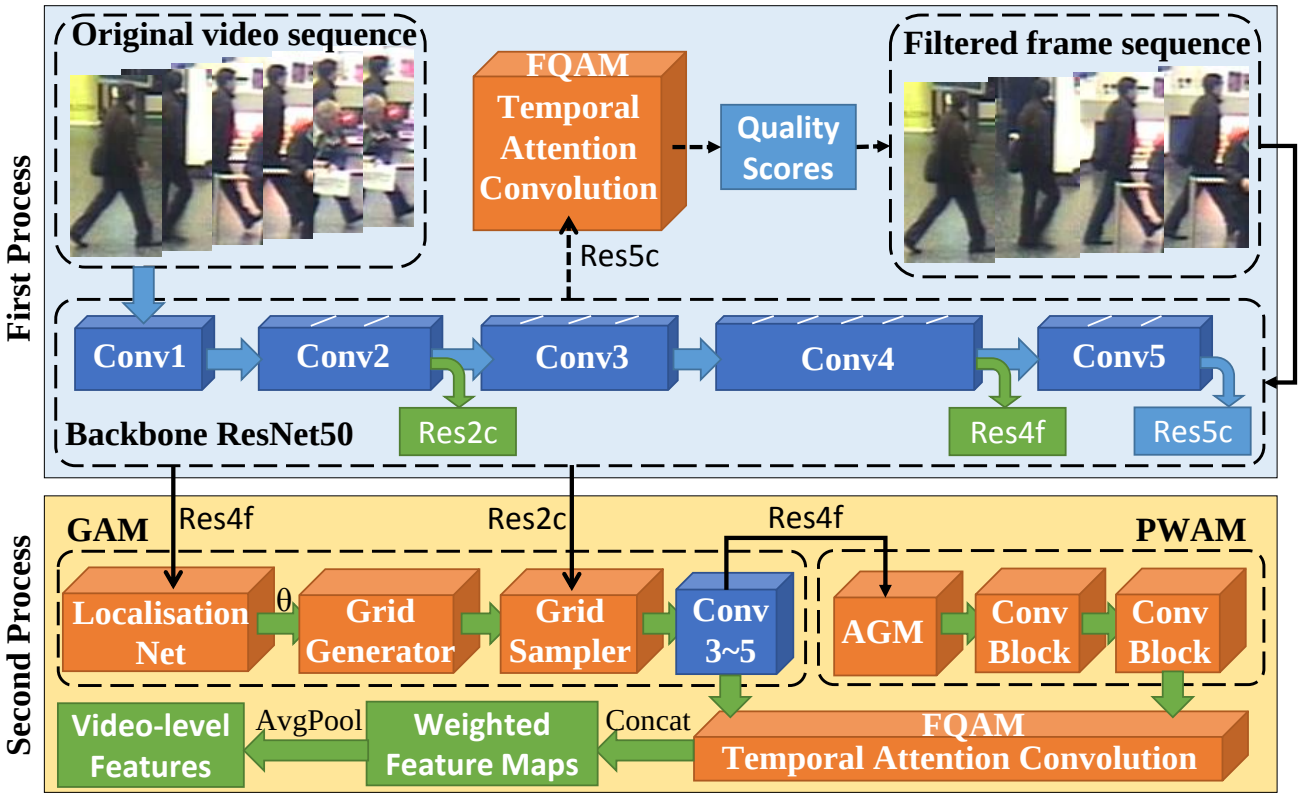


Figure 2. Illustration of the structure and workflow of the proposed FCFNet. It possesses three key sub-modules (i.e., PWAM, GAM and FQAM). The FCFNet first pre-trains the backbone, generating middle-level features and frame importance scores. Different sub-modules are embedded parallelly in different locations of the backbone.

3.2. Correcting large variances from image features

We design PWAM and GAM for reducing the background noise influence and aligning person areas, respectively, and the two sub-modules act on the model simultaneously to correct frame-level features in parallel.

3.2.1. Pixel-Wised attention module

In this parallel sub-module, we employ an attention-based method to reduce the attention resources allocated to the image noise parts under the convolutional receptive field, eliminate interference such as occlusions and hybrid backgrounds, and propose the Attention Generate Module (AGM). As shown in Figure 2, AGM is followed by two "Conv Block" connected by MaxPool to output the final features, where "Conv Block" is composed of a convolutional layer, a Batch Normalization (BN) layer and a ReLU activation layer.

The AGM presented in Figure 3 fetches the middle-level features "Res4f" of "Conv3 5" (the structure of the third to last stages in ResNet-50) as its input and outputs the predicted pixel-level importance score weighted features, where the importance score represents the importance of each pixel. The person area pixels that contribute to pedestrian identification will get a higher score, while the pixels of the obscured parts and the background areas will of course be given lower scores. After weighting by importance scores, the degree of influence of noise pixels on the representation of image features is mitigated. The AGM is a lightweight architecture composed of three convolution layers, the kernel size of the first convolution layer is 1×1 and its output dimension is 512, the kernel size of the second one is 3×3 and its output dimension is maintained at 512, the last convolution layer outputs the features of 1 channel with a kernel of 1×1. The activation function in conjunction with the first two convolution layers is ReLU. For the final output scores need to be rescaled to the [0,1] range using the Sigmoid function. Then, before multiplying with the original feature map, it needs to be expanded to the corresponding dimension. The above process of weighting the internal coordinates of the feature maps can be described as a concise set of formulas, in which the middle-level feature maps $F \in \mathbb{R}^{C \times H \times W}$ is used as the input of AGM, and the resulting score map can be represented as:

$$S = \sigma(f_{AGM}(F)) \in [0, 1], \tag{3.1}$$

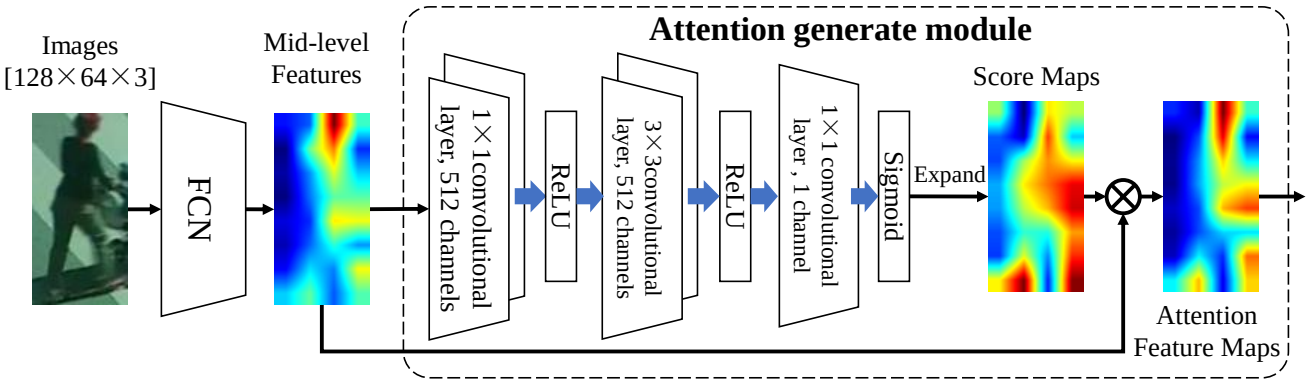


Figure 3. Structure of the Attention Generate Module. AGM takes the middle-level features as input. The probability map is generated through some convolution layers. Then each channel of the middle-level feature maps multiplies the score map to get the attention feature maps. Color maps are used here for viewing.

The Sigmoid function represented by $\sigma(\cdot)$ outputs score map S with dimension $H \times W$. And the

homologous attention feature map F_{att} can be obtained according to the following formula:

$$F_{att}(c, x, y) = S(x, y) \cdot F(c, x, y), \ c \in \{1, 2, ..., C\}. \tag{3.2}$$

The coordinate (x, y) is used to traverse all locations of the feature space. The illustration of AGM can be seen in Figure 3.

3.2.2. Global alignment module

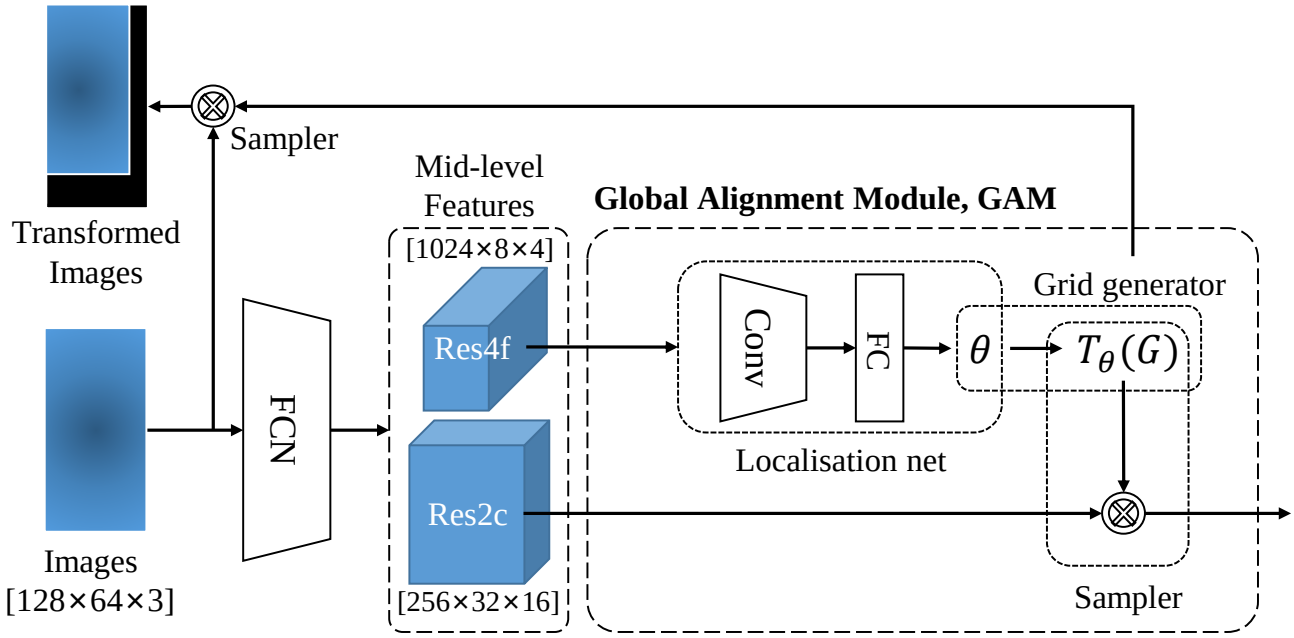


Figure 4. The details of the GAM. GAM takes the middle-level feature maps as input and learns the affine transformation matrix θ, outputs the transformed feature maps.

Due to the automatic detection of human areas and uncontrolled body movements, misalignment is a common problem in person re-id. STN [43] is a popular module for spatial transformation. It has shown effectiveness in image classification tasks. Inspired by this insightful work, we design the GAM shown in Figure 4. Here STN utilizes the global feature maps as the input to learn the affine transformation matrix. The STN consisted of localisation network, grid generator and sampler. The localisation network contains some hidden layers, which can take the middle-level feature maps "Res4f" as the input and outputs the transformation matrix θ. Output θ is used for spatial transformation and different transformation types correspond to different sizes of θ. In this paper, θ is a 2×3 matrix for the 2D affine transformation. The grid generator uses the affine transformation matrix θ as input to build a regular sampling grid $G = \{G_i\}$ at the "Res2c" scale, where $G_i = (x_i^t, y_i^t)$. The specific calculation process of affine transformation is represented as:

$$\begin{pmatrix} x_i^s \\ y_i^s \end{pmatrix} = T_\theta(G_i) = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix}. \tag{3.3}$$

where (x_i^s, y_i^s) is used to traverse the input feature maps, and (x_i^t, y_i^t) corresponds to the target coordinates mapped into the output features according to the regular grid. Finally, the sampler generates the sampled and corrected resulting features from the input feature "Res2c" according to the sampling grid. The necessity for rotation is often not obvious as the pedestrians tend to be standing postures in person re-id task. So we fix $\theta_{12} = \theta_{21} = 0$ in Eq (3.3) and learn the other four parameters (i.e. the parameters for translation and scale). As shown in Figure 2, the corrected features of the GAM continue to be transmitted to "Conv3 5" to output the final features.

3.3. Frame-level feature fusion and evaluation

We design the FQAM, and its workflow is shown in Figure 5, which introduces an inter-frame attention mechanism based on the convolutional layer to provide weights for frame-level feature fusion and importance scores for frame quality evaluation.

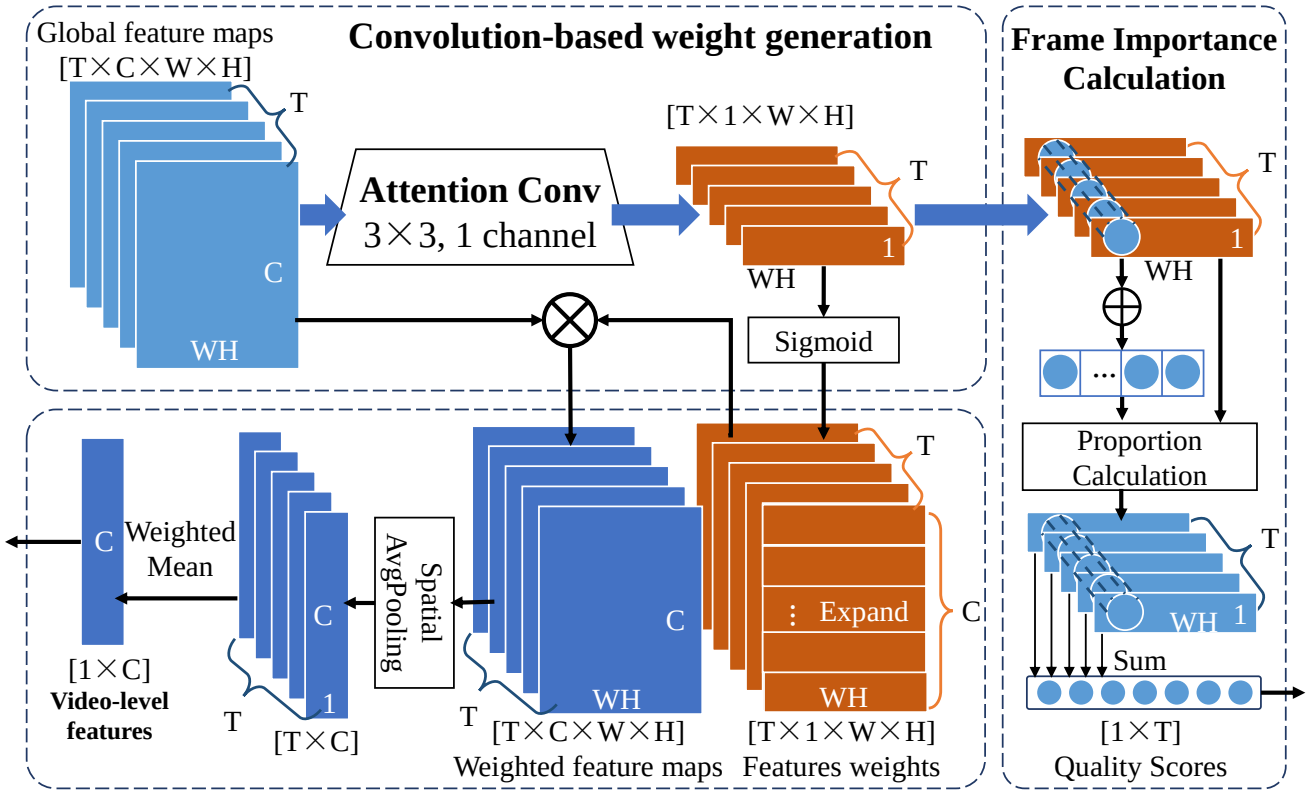


Figure 5. The workflow of the FQAM. This module introduces an inter-frame attention mechanism, which guides informative features to dominate the fusion process, and generates frame importance scores to optimize model input.

3.3.1. Convolution-based weights generation

The image is mapped to the global feature map $F \in \mathbb{R}^{C\times H\times W}$ after spatial feature extraction, and the frame image sequence is converted into a sequence consisting of multi-dimensional frame-level

feature maps. If temporal pooling is applied directly to the feature sequence, the values of each element on the frame feature will be integrated into the average or maximum values of the corresponding dimensions. However, the direct action between high and low activation values does not take into account the different frame quality and information content. The pixel-level attention mechanism focuses on differences in importance of pixels in an image. Similarly, we can introduce an inter-frame attention mechanism to adjust the component values at the level of serial data to enhance the comparability of corresponding points in different frame-level feature maps. Besides RNN and LSTM [19, 20, 44], some work uses convolutional networks to process serial data. The convolution layer is simple in structure, allows parallel computing output. Gao et al. [22] further validated the feasibility of generating weights by convolution. We employ a 3×3 convolution layer with 1 channel to model the elements that are already present in the feature map sequence to predict frame-level weights, as shown in Figure 5. The $T \times C \times W \times H$ multi-channel feature information is processed in parallel, and each frame feature's appropriate weights $V \in \mathbb{R}^{H \times W}$ for value adjustment is obtained using the nonlinearity of the convolution layer. After activation and expansion, $V \in \mathbb{R}^{C \times H \times W}$ is assigned to the original frame features. By using the loss function constraint, the quality information is incorporated in advance for later average fusion.

3.3.2. Frame importance scores calculation

Intuitively, the introduction of the frame-level attention mechanism takes the feature quality level of the entire frame sequence into account. The larger the generated weights, the greater the dominance of the corresponding elements in mean fusion, and the more information provided for the model to identify pedestrian identity. We believe that the uneven quality of features originates from the difference in image content. Video sequences captured in actual application scenarios contain more noise, misalignment and other bias interference. Some frames with poor quality even hinder the overall video-level feature representation. By introducing frame-level attention mechanism, the ability of the image to express pedestrian identity can be measured numerically. We can reuse the frame-level attention weights, filter the input of the model by deleting some low-quality frames, improve the quality of feature expression and reduce training overhead. As shown in Figure 5, the T frame-level features $F \in \mathbb{R}^{1 \times C}$ fuse with each other at the same element location, and the weight of one element on the feature map only affects the importance of elements between frames in the first dimension. Therefore, direct summation of the values on a whole weight map can easily result in the small weights corresponding to the elements with large values being overwritten by the large weights of the small value elements. For a video sequence, we regularized all the weight maps values $V \in \mathbb{R}^{H \times W}$ to the equivalent ratio level, i.e. under different feature maps, the elements' weights in the corresponding locations are converted into percentages and then sum them as frame quality scores (Qs), as described below.

$$v_{x,y}^{i} = \frac{v_{x,y}^{i}}{\sum_{i=0}^{T-1} v_{x,y}^{i}} \times 100.0, i \in [0, T-1], x \in [0, W-1], y \in [0, H-1] \tag{3.4}$$

$$Qs_i = \sum_{x=0}^{W-1} \sum_{y=0}^{H-1} v_{x,y}^{i} \tag{3.5}$$

where i traverses the T-frame and (x, y) traverses the elements of the two-dimensional weights $V \in \mathbb{R}^{H\times W}$, and $v^i_{x,y}$ is the weight value after conversion to ratio at position (x, y).

4. Experiments

In the experimental session, to appraise the proposed FCFNet, we implement extensive experiments based on four popular datasets [3–5,45]. We started with the ablation studies to examine the rationality of each component in our framework. Then, we present the overall results and contrast the proposed network with some recent methods to evaluate its effectiveness. All experiments are based on the NVIDIA GeForce GTX 1080 Ti GPU and implemented using PyTorch.

(a) MARS **(b)** DukeMTMC-video

(c) PRID **(d)** ILIDS-VID

Figure 6. Sample frames from four benchmark datasets. Randomly select one video in each dataset, and then extract 6 frames from each video.

4.1. Dataset and evaluation metric

We experiment with four benchmark datasets, which are widely used in many works. As shown in Figure 6, the large variance makes these datasets challenging, especially in iLIDS-VID [4], which includes frequent misalignments, occlusions, and other noise. **MARS** [3] is a large-scale video-based re-id dataset, its trajectory sequences are automatically generated by DPM and GMMCP [46]. The detection and tracking errors make the records contain more interference. This dataset contains 1261 pedestrian trajectory video sequences captured by 6 cameras with an average of 60 frames. There are 625 persons for training, corresponding to 8298 videos. The remaining 636 persons are used as a test set, of which 1980 sequences are for queries and 9330 are for the gallery. **DukeMTMC-video** [45] uses DPM target detection, involving 8 cameras, comprising 1812 pedestrians and 4832 video track sequences containing 169 frames on average. 2196 videos of 702 persons are used for training, 702 videos of the other 702 persons for querying, and 1934 videos of 1110 persons in the gallery, of which 408 are interferers. **PRID** [5] is a handcrafted dataset collected by 2 static cameras in a spacious outdoor environment with less background noise and obstruction. The 200 pedestrians in the dataset have 1 record in each of the two perspectives, totaling 400 videos, averaging 100 frames. The training

set randomly selected 100 persons, and the other 100 persons became the test set. The training set uses a total of 200 video sequences under both cameras concurrently, while the query set and the gallery set are from the remaining 100 sequences under two cameras respectively. **ILIDS-VID** [4] is taken from the airport reception hall during busy hours. The video tracks of 300 pedestrians are captured by 2 cameras without overlapping fields of vision. Each pedestrian has a pair of handcrafted tracks, including 600 videos, with an average of 71 frames. 300 videos of 150 persons are used as a training set, and the videos of the other 150 people under two cameras serve as querying and gallery, respectively.

We utilize the standard evaluation protocol to evaluate model recognition ability, on all benchmark datasets, we use the Cumulative Matching Characteristics (CMC) at rank-1, rank-5 and rank-10 to evaluate the ranking hitting accuracy of model retrieval targets, and utilize the mean Average Precision (mAP) to evaluate the overall retrieval effect of the model.

4.2. Implementation details

4.2.1. Backbone

In some person re-id studies [6, 47, 48], the IDE model [49] is a popular deep learning baseline, which uses ResNet-50 [42] as the backbone. When IDE model is applied to some particular tasks, the output dimension of its last fully-connected layer is often set to the identities' number of the input set (e.g. 625-dim in the MARS training set). The parameters pre-trained in ImageNet [50] will provide support for model initialization. During testing, the outputs of the last pooling layer "pool5" of the entire backbone are fused as the frame-level feature representations. Inspired by the effectiveness of BNNeck [51], the last stride of our ResNet-50 is 1. We use the strategy of direct addition of cross entropy ID loss and triple loss to learn person features, so that the original "pool5" layer features are used for triplet loss calculation, and then attach a BN layer to the "pool5" layer to restrict features to the supersphere, reduce the feature distribution area, and use the features behind the BN layer to accelerate the convergence of ID loss. Under this structure, we take features passing through the BN layer as the last frame-level features.

4.2.2. Training strategy

In order to implement batch gradient descent, the following video frame sampling methods are used to process video sequences with different lengths. For the training process, we randomly extract T = 6 frames from each video sequence as input clips. After several epochs, most frame images in the video have the opportunity to participate in the training.

We set the training batch size to 24, for larger MARS [3] and DukeMTMC-video [45] datasets with more sampling cameras, we extract 4 clips for each of the 6 identities in the same batch. For smaller PRID [5] and iLIDS-VID [4] datasets, we extract 2 clips for each of the 12 identities. The model iterates for 200 epochs with parameters tuned by the Adam optimizer, which we set to have an initial learning rate of 1.0×10^{-3} and a weight decay of 5.0×10^{-4}. Meanwhile, we adopt a warm-up strategy to promote more stable model convergence, specifically by linearly incrementing the learning rate from a smaller value of 3.5×10^{-6} to a preset initial learning rate of 3.5×10^{-4} during the first 10 epochs of training, while adopting the MultiStepLR strategy, the learning rate decayed by 0.3 every 35 epochs. We resize all images to 128×64 pixels, smaller than most video-based person re-id studies. In

order to enrich the input data during the training phase, we introduce random erasing [52] and random horizontal flipping strategies, setting both their execution probabilities to 0.5, and finally input the normalized frame images into the model.

4.2.3. Testing details

In testing, for queried and gallery videos, we divide the whole video sequence into T = 6 consecutive clips, each clip extracts video-level features separately, and then all video-level features are averaged to serve as the ultimate feature of the entire video. If the fragment is not long enough, fill it with the existing frame image.

Based on the above backbone structure, the baseline uses the output features adjusted by the BN layer as the frame-level feature representation. Based on feature fusion, the GAM module acts on the front end of the backbone without affecting the number of final output branches. The size of model output features is the same as the baseline (i.e., 2048 dim), while the PWAM module acts on the back end of the backbone, adding another output branch, we test the independent PWAM branch (i.e., 2048 dim) and its combination with the backbone output (i.e., 2048 $\times$ 2 dim).

4.3. Ablation study

In this subsection, following the same experimental settings, we conduct several equitable self-reflection experiments on each sub-module based on PRID and iLIDS-VID to certify the design rationality and effectiveness of FQAM, PWAM and GAM components of the FCFNet, respectively.

Table 1. Ablation study results of feature fusion methods. Rank-1,5,10 Accuracy (%) and mAP (%) are shown.

Fusion Methods	PRID				ILIDS-VID			
	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP
Temporal MaxPooling	85.00	96.00	99.00	89.98	72.67	90.00	97.33	80.67
Temporal AvgPooling	85.00	96.00	99.00	90.11	72.00	91.33	94.67	80.68
Feature Attention	89.00	96.00	98.00	92.18	78.67	90.00	94.00	84.15

4.3.1. Effectiveness of the FQAM

We design the Frame Quality Assessment Module (FQAM) to support the last frame-level feature fusion and input frames filtering. We will design two-part ablation experiments to verify the advantages of FQAM structure design, show the effect of filtering frames under each dataset, and determine the appropriate deletion proportion.

First, to verify that the convolution-based feature weight generation of FQAM is better than ordinary temporal pooling, we perform feature fusion methods effect contrasts as shown in Table 1 without removing any frames. As shown in the third line, the attention mechanism was used to adjust the strength of inter-frame feature interaction intensity, and to facilitate feature weighting fusion into video-level features. On PRID, 89.00% of rank-1 and 92.18% of mAP are obtained, and on iILIDS-VID, the rank-1 and mAP are 78.67 and 84.15%, respectively. Compared with the other two temporal pooling methods, feature attention has achieved significant improvement. On PRID, rank-1 and MAP increased by 4.00 and 2.20% respectively compared with MaxPooling. And compared with

AVGPooling, rank-1 and mAP increased by 4.00 and 2.07%. On iILIDS-VID, compared with the two pooling methods, rank-1 increased by 6.00 and 6.67% respectively, and mAP increased by 3.48 and 3.47% respectively. The first two pooling methods have similar results. MaxPooling can cause severe information loss by discarding all the non-maximum activation values in most features. And AvgPooling may counteract the high and low activation values and lose discriminant information. The feature attention adjusts the values of features so that the corresponding features of the information frames can play more roles, enhancing the comparability between the frame-level features.

The second part of the ablation experiment in this module is to find the appropriate filtering proportion on each dataset. After experimentation as shown in Table 2, we find that deleting some frames can even improve the model detection capability compared to the baseline model of keeping all frames, implying that by eliminating some low-quality frames, the model can adapt to the overall video quality degradation caused by the reduction of pixels. Deleting some poor frames can reduce the overhead of model training, speed up detection, and reduce the storage of redundant information.

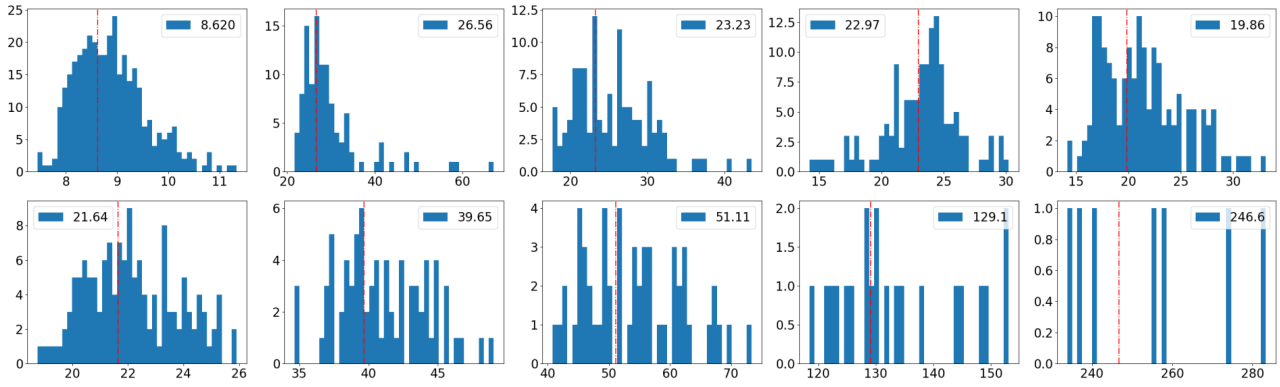
Table 2. Proportion validation of interference frames of each dataset. Rank-1 Accuracy (%) and mAP (%) Are Shown. The underlined numbers correspond to the outcome at the final chosen proportion.

Proportion	MARS		DukeMTMC-video		PRID		ILIDS-VID	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
0.00	79.40	65.06	88.75	87.05	89.00	92.18	78.67	84.15
0.05	78.91	64.28	88.75	86.95	88.00	91.66	78.67	84.33
0.10	78.91	64.45	89.03	86.83	89.00	92.68	77.33	84.41
0.15	79.18	64.87	89.46	87.51	86.00	91.49	78.67	84.80
0.20	78.53	64.06	89.89	86.91	86.00	91.32	<u>81.33</u>	<u>87.26</u>
0.25	78.64	64.52	89.60	87.16	88.00	92.65	74.00	81.74
0.30	78.26	63.77	88.89	87.45	86.00	90.74	74.00	81.82
0.40	<u>79.08</u>	<u>64.12</u>	89.32	87.23	<u>90.00</u>	<u>93.37</u>	72.67	81.85
0.50	77.39	63.14	<u>89.32</u>	<u>87.36</u>	89.00	92.67	66.00	75.77
0.60	78.21	62.01	89.46	87.09	88.00	92.00	62.00	70.35

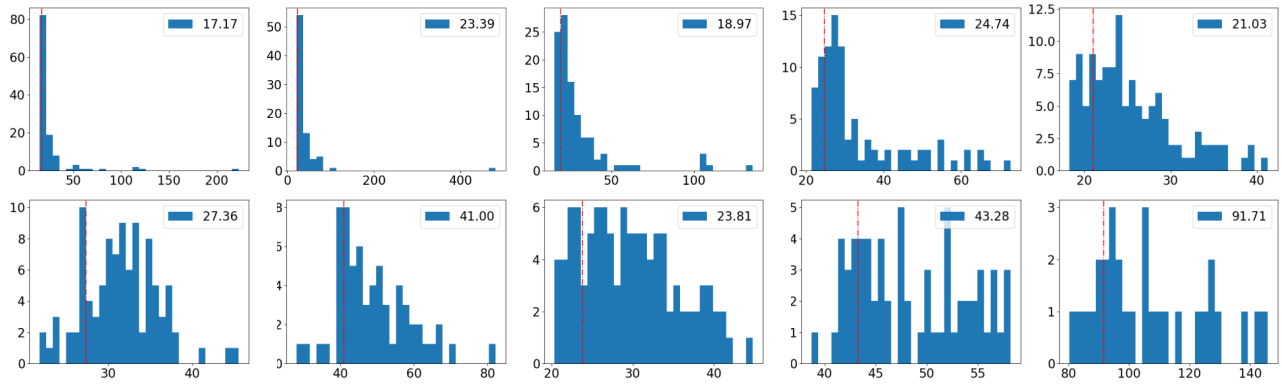
As shown in Table 2, when all generated scores are ranked in ascending order, we tested various outcomes after deleting frames of distinct proportions. According to the experimental results, we finally determined the proportion of subsequently deleted frames on the four datasets to 40, 50, 40 and 20%, respectively. On MARS, we actually try to swap a slight decline in indicators for a noticeable lightweight of the dataset, with only 0.32 and 0.94% gaps from the 79.40% rank-1 and 65.06% mAP of the baseline. On DukeMTMC-video, removing half of the frames still maintains good detection accuracy. On PRID, when 40% of the low-score images are deleted, a certain degree of improvement can be achieved, rank-1 and mAP increase by 1.00 and 1.19%, respectively. On iLIDS-VID, when 20% of the frames are deleted, they increase by 2.66 and 3.11%, respectively. As the deletion proportion increases gradually, the results are close to or even slightly lower than the baseline. Take PRID as an example, when 15, 20 and 30% are deleted, rank-1 decreases by 3.00% and mAP decreases by 0.69, 0.86 and 1.44%, respectively. After only some of the low-quality frames are deleted, the interference of the remaining inferior frames cooperating with the deleted frames on the

overall feature expression will become more obvious. Therefore, significant growth can be achieved after the appropriate proportion has been achieved, i.e., interference items have been completely removed. On PRID, 10% more deletion than 30% can increase the rank-1 and mAP by 4.00 and 2.63% respectively. However, more deletions will affect the normal expression of frame features, especially on iLIDS-VID. As the deletion proportion increases, rank-1 decreases rapidly.

The above data results can also be complemented by Figure 7, which presents the quality score distributions of video frames of various lengths under each dataset. The more frames there are, the smaller the values scattered across the weight maps and the smaller the sum. The red threshold corresponding to the last chosen deletion proportion is often located at the local minimum of the score distribution histogram. When there are fewer deletions, low-quality frames with no supplementary information still exist. There is a high probability that the number of frames near the deletion threshold will be large. As the deletion proportion increases, a great amount of information will be lost. It is evident in the iLIDS-VID corresponding to the bottom half of Figure 7 that the frame scores near the red deletion threshold line are densely and uniformly distributed, and a slight increase in the proportion may result in the loss of plenty of valid frames.



(a) PRID



(b) ILIDS-VID

Figure 7. Quality score distributions of video sequences: (a) PRID, (b) ILIDS-VID. Each subgraph corresponds to a video, where the horizontal and vertical axes correspond to the frame quality score and the number of frames falling into the interval, respectively. The red line identifies the screening score and the corresponding value is marked above.

4.3.2. Effectiveness of the GAM

With model parameters continuously updated iteratively to find the appropriate affine transformation matrix θ, the GAM enables models to shift, zoom and crop the images with more background areas and fills the missing parts with zero value to increase the proportion of pedestrian information in the whole feature construction process. As shown in Table 3, inserting GAM directly onto the baseline (+ GAM), rank-1, 5, 10 and mAP increased by 1.00, 2.00, 2.00 and 1.65% on PRID, respectively. On iLIDS-VID, the increases are 4.00, 5.33, 2.67 and 3.99%, respectively. The iLIDS-VID is collected in busy public places. The background of its video images is much more complex than PRID. Removing some complicated background information has a more obvious improvement. To verify the desirability of the learning of transformation matrix θ, we fix the θ in Eq 3.3, and no affine transformation is implemented on the input feature maps (GAM/Fixθ). In this case, the rank-1 and mAP of GAM on PRID decreased by 2.00 and 1.47%, respectively, and on iLIDS-VID decreased by 2.67 and 1.54%. We can notice that the introduction of STN improves the recognition ability of the model in both datasets.

Table 3. Ablation study results of PWAM and GAM on PRID and iLIDS-VID. Rank-1,5,10 accuracy (%) and mAP (%) are shown.

Models	PRID				ILIDS-VID			
	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP
Baseline	89.00	96.00	98.00	92.18	78.67	90.00	94.00	84.15
+ PWAM1/FixS	87.00	96.00	99.00	91.25	78.67	92.67	95.33	85.36
+ PWAM1	90.00	95.00	98.00	92.68	80.00	96.67	98.00	87.09
+ PWAM2	91.00	98.00	100.00	94.04	81.33	94.00	96.67	86.82
+ GAM/Fixθ	88.00	99.00	99.00	92.36	80.00	94.67	98.00	86.60
+ GAM	90.00	98.00	100.00	93.83	82.67	95.33	96.67	88.14
FCFNet	93.00	97.00	100.00	94.83	82.67	96.67	96.67	88.59

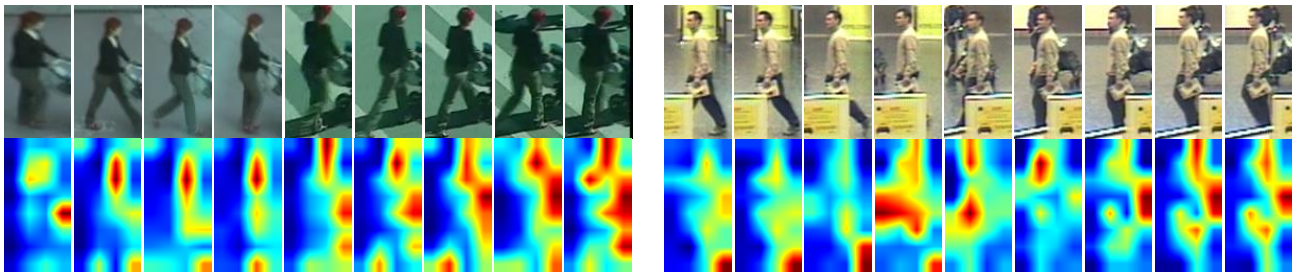
4.3.3. Effectiveness of the PWAM

As shown in Figure 3, the pixel-level attention mechanism introduced by PWAM coordinates the importance of each pixel within a middle-level feature. The importance score generated by the AGM enables the model to focus more on the pixels that contribute to person identification. In this part of the ablation experiments, we mainly verify the necessity of generating pixel scores and test two schemes of PWAM matching the output of the backbone.

First, as shown in Table 3, when the PWAM branch is inserted directly onto the baseline backbone and only the output of PWAM is used to identify pedestrians (+ PWAM1), the rank-1 and mAP of PRID increase by 1.00 and 0.50%, and the rank-1 and mAP of iLIDS-VID increase by 1.33 and 2.94%, respectively, compared with the baseline. The improvement is more obvious on the iLIDS-VID with serious background noise. To verify that PWAM plays an active role in the FCFNet, we eliminate the adjustment effect of the score maps by fixing the corresponding score for each coordinate point to 1 (+ PWAM1/FixS). As can be seen in Table 3, the mAP improved by 1.43 and 1.73% when the score maps are updated properly. This indicates that the PWAM can comprehend how to adjust the significance of each pixel from the input feature maps, and reduce image noise interference by focusing on pedestrian areas.

Next, the PWAM is inserted into the second-to-last stage of the backbone, adding a new output branch to the entire network structure. We tested the effects of the PWAM branch working alone (+ PWAM1) and direct concatenating with the output of original backbone (+ PWAM2). Merging the two branches (+ PWAM2) also has a slight advantage, doubling the feature dimension to contain more information. Compared with the baseline, rank-1,5,10 of PRID can be increased by 2.00% and mAP by 2.79%. Meanwhile, rank-1,5,10 and mAP of iLIDS-VID are increased by 2.66, 4.00, 2.67 and 2.67%, respectively.

Figure 8 further shows the visualization of weighted attention feature maps under two video sequences, which indicates that the feature maps pay more attention to the human body regions, and ignore image noise and occlusions.



(a) PRID **(b)** ILIDS-VID

Figure 8. Feature Visualization with Pixel Level Attention Mechanism. The first row of each group is the input native frames, and the next row is the attention feature maps of the AGM module. Best viewed in color.

4.4. Experimental results

The performance of our modified IDE baseline is reported in the first column of Table 4. The rank-1,5,10 accuracy and mAP on MARS are 79.40, 91.41, 93.42 and 65.06%, the corresponding evaluation indices on DukeMTMC-video are 88.75, 98.72, 99.43 and 87.05%, respectively. 89.00, 96.00, 98.00 and 92.18% on the PRID and 78.67, 90.00, 94.00 and 84.15% on iLIDS-VID, respectively. Table 2 and Table 3 illustrate that the three sub-modules we employed all ameliorate the results compared with the baseline, which also demonstrates the effectiveness of each sub-module. And the appropriate deletion proportion under each dataset was also determined. The second column (FQAM) shown in Table 4 shows the results of deleting 40, 50, 40 and 20% of the four datasets, respectively. In this subsection, we further test the effect of inserting two other sub-modules based on FQAM. Compared with the baseline, when FQAM and PWAM simultaneously act on the backbone (FQAM + PWAM), rank-1 on the four datasets increases by 1.80, 2.13, 4.00 and 3.33%, respectively, and mAP increases by 2.78, 2.27, 2.78 and 3.20%, respectively. When both FQAM and GAM are applied to the backbone (FQAM + GAM), rank-1 indicators are 1.80, 2.28, 3.00 and 4.66% higher than the baseline, and mAP indicators are improved by 2.26, 2.01, 1.96, and 3.78%, respectively.

Table 4. The effects of different sub-modules and last FCFNet on four datasets. Rank-1,5,10 accuracy (%) and mAP (%) are shown. The top results are shown in boldface.

Models		Baseline	FQAM	FQAM + PWAM	FQAM + GAM	FCFNet	Weighted Stitching
Feature dimension		2048×1	2048×1	2048×2	2048×1	2048×2	2048×3
MARS	Rank-1	79.40	79.08	81.20	81.20	**81.41**	**81.14**
	Rank-5	91.41	90.65	92.01	91.68	**92.66**	91.79
	Rank-10	93.42	93.64	94.46	93.97	**95.05**	93.97
	mAP	65.06	64.12	67.84	67.32	**69.24**	67.50
DukeMT MC-video	Rank-1	88.75	89.32	90.88	91.03	**91.03**	**91.03**
	Rank-5	98.72	98.29	98.86	98.72	**98.86**	98.72
	Rank-10	99.43	99.00	99.29	99.43	**99.57**	99.43
	mAP	87.05	87.36	89.32	89.06	**89.77**	89.09
PRID	Rank-1	89.00	90.00	93.00	92.00	**93.00**	**93.00**
	Rank-5	96.00	98.00	98.00	96.00	97.00	**98.00**
	Rank-10	98.00	100.00	98.00	98.00	**100.00**	99.00
	mAP	92.18	93.37	94.96	94.14	94.83	**95.32**
ILIDS-VID	Rank-1	78.67	81.33	82.00	83.33	82.67	**83.33**
	Rank-5	90.00	94.67	95.33	94.00	**96.67**	94.67
	Rank-10	94.00	98.00	97.33	96.67	96.67	**98.00**
	mAP	84.15	87.26	87.35	87.93	**88.92**	88.39

Our FCFNet has two feature correction sub-modules inserted respectively in the front and rear of the backbone, and guides frame-level feature fusion with FQAM to achieve synergies between sub-modules. On MARS, the rank-1,5,10 and mAP of FCFNet reached 81.41, 92.66, 95.0 and 69.24%, respectively. Indicators increased by 2.01, 1.25, 1.63 and 4.18% compared to baseline. On DukeMTMC-video, rank-1 and mAP are elevated by 2.28 and 2.72%. The mAP improved by 2.65 and 4.77% on PRID and iLIDS-VID respectively.

FCFNet follows the dual branch combination strategy of PWAM, so the output feature dimension is 2048×2. FCFNet applies each sub-module to different stages of the backbone, to verify the reliability of this modular combination strategy, we additionally test the effect of output feature weighted stitching of FQAM + PWAM and FQAM + GAM. The corresponding weights for stitching here are the degrees of improvement of each model compared with the baseline on mAP. Obviously, as shown in the last column of Table 4, features weighted concatenated into 2048×3 dimensions will contain more information and may also achieve better results. The modular combination of FCFNet realizes information aggregation, reduces feature dimensions relative to 2048×3, and each indicator is comparable to feature stitching.

4.5. Comparisons with the recent methods

Most popular video-based person re-id methods are naturally based on the original 256×128 images, but our method is mainly aimed at the small-size person images, i.e., 128×64, which reduces the amount of available pixel information and increases the challenge. As shown in Table 5, to

compare with the methods in recent years, we also test the FCFNet in the 256 × 128 configuration.

Under 256 × 128 resolution, the GCN-based methods MGH [59] and AdaptiveGraph [58], as well as STMN [44] and BiCnet-TKS [16], which focus on spatial and temporal dimension information, show good performances. In these two categories, we test the re-id effect of MGH and STMN at 128 × 64. Compared with MGH, rank-1 and rank-5 on PRID can achieve 3.0 and 4.5% improvement, 6.0 and 12.7% higher on iLIDS-VID with more noise, indicating that FCFNet can discriminate valid person information from fewer pixels.

Table 5. Comparisons with recent methods on mainstream datasets.

Input Size	Methods	MARS		PRID		ILIDS-VID	
		Rank-1	mAP	Rank-1	Rank-5	Rank-1	Rank-5
256 × 128	SFT [20]	70.6	50.7	79.4	94.4	55.2	86.5
	CDIN [53]	72.5	53.6	86.6	95.9	71.2	90.2
	Scale-fusion [54]	75.4	69.3	84.2	95.8	75.1	92.8
	HIRF [55]	80.1	63.4	81.0	91.8	65.4	83.1
	CSACSE [25]	81.2	69.4	88.6	99.1	79.8	91.8
	MSTA [56]	82.3	69.4	87.6	96.0	65.3	85.0
	E-GLRN [32]	83.3	70.8	91.6	99.7	81.3	93.5
	M3D [40]	84.4	74.1	94.4	**100.0**	74.0	94.3
	CSACSE+OF [25]	86.3	76.1	93.0	99.3	85.4	**96.7**
	FGRA [57]	87.3	81.2	95.5	**100.0**	**88.0**	**96.7**
	AdaptiveGraph [58]	89.5	81.9	94.6	99.1	84.5	**96.7**
	BiCnet-TKS [16]	90.2	**86.0**	–	–	75.1	84.6
	MGH [59]	90.0	85.8	94.8	99.3	85.6	97.1
	STMN [44]	**90.5**	84.5	95.0	97.0	86.0	90.8
	FCFNet	88.0	82.6	**96.0**	99.0	87.3	**96.7**
128 × 64	MGH [59]	78.2	69.7	90.0	92.5	76.7	84.0
	STMN [44]	80.5	**69.2**	89.0	92.3	72.7	80.6
	FCFNet	**81.4**	**69.2**	**93.0**	**97.0**	**82.7**	**96.7**

5. Conclusions

In some public surveillance videos, the pixel sizes of person areas tend to be closer to 128 × 64, less pixel information leads to the overall degradation of the video frame quality, and it is harder to distinguish the pedestrian information from interferences. We construct a Person Feature Correction and Fusion Network (FCFNet) with three well-designed sub-modules to extract the features contained in small-sized images by combining two perspectives: correcting the spatial features and fusing more high-quality features. After multilayer convolution, the entanglement between person information and unrelated information becomes tighter. For the complex variance in pedestrian recognition, we consider using affine transformation in GAM to eliminate most of the redundant background at the front of the model input. Considering that person area is irregular due to noise such as pedestrian posture change and occlusions, we further design PWAM, which combines pixel-level attention mechanism and utilizes semantic information from deeper layers to refine pedestrian features. When the frame quality

decreases, some frames may not help to form discriminative features. Based on the attentional weighted feature fusion, we design the FQAM to obtain robust video-level features and filter the datasets through the frame quality scores. Filtering the frame images while ensuring accuracy can reduce training costs and improve detection speed. The experimental results demonstrate that our FCFNet can enhance the recognition performance on four mainstream datasets under smaller image sizes 128×64.

Acknowledgments

This work was partly supported by the National Natural Science Foundation of China (Grant No. 72274058).

Conflict of interest

The authors declare no conflict of interest.

References

1. M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, S. C. Hoi, Deep learning for person re-identification: A survey and outlook, *IEEE Trans. Pattern Anal. Mach. Intell.*, **44** (2021), 2872–2893. https://doi.org/10.1109/TPAMI.2021.3054775
2. Z. Ming, M. Zhu, X. Wang, J. Zhu, J. Cheng, C. Gao, et al., Deep learning-based person re-identification methods: A survey and outlook of recent works, *Image Vision Comput.*, **119** (2022), 104394. https://doi.org/10.1016/j.imavis.2022.104394
3. L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, et al., Mars: A video benchmark for large-scale person re-identification, in *European conference on computer vision*, Springer, (2016), 868–884.
4. T. Wang, S. Gong, X. Zhu, S. Wang, Person re-identification by video ranking, in *European conference on computer vision*, Springer, (2014), 688–703.
5. M. Hirzer, C. Beleznai, P. M. Roth, H. Bischof, Person re-identification by descriptive and discriminative classification, in *Scandinavian conference on Image analysis*, Springer, (2011), 91–102.
6. Z. Zheng, L. Zheng, Y. Yang, Pedestrian alignment network for large-scale person re-identification, *IEEE Trans. Circuits Syst. Video Technol.*, **29** (2019), 3037–3045. https://doi.org/10.1109/TCSVT.2018.2873599
7. C. Su, J. Li, S. Zhang, J. Xing, W. Gao, Q. Tian, Pose-driven deep convolutional model for person re-identification, in *2017 IEEE International Conference on Computer Vision*, (2017), 3980–3989. https://doi.org/10.1109/ICCV.2017.427
8. C. Wang, Q. Zhang, C. Huang, W. Liu, X. Wang, Mancs: A multi-task attentional network with curriculum sampling for person re-identification, in *European Conference on Computer Vision*, 2018. https://doi.org/10.1007/978-3-030-01225-0_23
9. Y. Liu, J. Yan, W. Ouyang, Quality aware network for set to set recognition, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, (2017), 5790–5799.

10. C. Chen, M. Qi, G. Huang, J. Wu, J. Jiang, X. Li, Learning discriminative features with a dual-constrained guided network for video-based person re-identification, *Multimed. Tools Appl.*,**80** (2021), 28673–28696. https://doi.org/10.1007/s11042-021-11072-y

11. S. Wang, B. Huang, H. Li, G. Qi, D. Tao, Z. Yu, Key point-aware occlusion suppression and semantic alignment for occluded person re-identification, *Inform. Sci.*, **606** (2022), 669–687. https://doi.org/10.1016/j.ins.2022.05.077

12. Z. Zhu, Y. Luo, S. Chen, G. Qi, N. Mazur, C. Zhong, et al., Camera style transformation with preserved self-similarity and domain-dissimilarity in unsupervised person re-identification, *J. Vis. Commun. Image Represent.*, **80** (2021), 103303. https://doi.org/10.1016/j.jvcir.2021.103303

13. S. Li, F. Li, K. Wang, G. Qi, H. Li, Mutual prediction learning and mixed viewpoints for unsupervised-domain adaptation person re-identification on blockchain, *Simul. Model Pract. Theory*, **119** (2022), 102568. https://doi.org/10.1016/j.simpat.2022.102568

14. H. Li, N. Dong, Z. Yu, D. Tao, G. Qi, Triple adversarial learning and multi-view imaginative reasoning for unsupervised domain adaptation person re-identification, *IEEE Trans. Circuits Syst. Video Technol.*, **32** (2021), 2814–2830. https://doi.org/10.1109/TCSVT.2021.3099943

15. Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Z. Hu, C. Yan, et al., Improving person re-identification by attribute and identity learning, *Pattern Recognit.*, **95** (2019), 151–161. https://doi.org/10.1016/j.patcog.2019.06.006

16. R. Hou, H. Chang, B. Ma, R. Huang, S. Shan, Bicnet-tks: Learning efficient spatial-temporal representation for video person re-identification, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2021), 2014–2023. https://doi.org/10.1109/CVPR46437.2021.00205

17. H. Li, Y. Chen, D. Tao, Z. Yu, G. Qi, Attribute-aligned domain-invariant feature learning for unsupervised domain adaptation person re-identification, *IEEE Trans. Inform. Forensics and Secur.*, **16** (2020), 1480–1494. https://doi.org/10.1109/TIFS.2020.3036800

18. Y. Wang, G. Qi, S. Li, Y. Chai, H. Li, Body part-level domain alignment for domain-adaptive person re-identification with transformer framework, *IEEE Trans. Inform. Forensics and Secur.*, **17** (2022), 3321–3334. https://doi.org/10.1109/TIFS.2022.3207893

19. 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), 1325–1334. https://doi.org/10.1109/CVPR.2016.148

20. Z. Zhou, Y. Huang, W. Wang, L. Wang, T. Tan, See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, (2017), 4747–4756. https://doi.org/10.1109/CVPR.2017.717

21. Y. Yan, B. Ni, Z. Song, C. Ma, Y. Yan, X. Yang, Person re-identification via recurrent feature aggregation, in *European conference on computer vision*, Springer, (2016), 701–716.

22. J. Gao, R. Nevatia, Revisiting temporal modeling for video-based person reid, preprint, arXiv:1805.02104.

23. Y. Zhao, X. Shen, Z. Jin, H. Lu, X. S. Hua, Attribute-driven feature disentangling and temporal aggregation for video person re-identification, in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, (2019), 4913–4922. https://doi.org/10.1109/CVPR.2019.00505

24. T. Rahman, M. Rochan, Y. Wang, Convolutional temporal attention model for video-based person re-identification, in *2019 IEEE International Conference on Multimedia and Expo (ICME)*, (2019), 1102–1107. https://doi.org/10.1109/ICME.2019.00193

25. D. Chen, H. Li, T. Xiao, S. Yi, X. Wang, Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, (2018), 1169–1178. https://doi.org/10.1109/CVPR.2018.00128

26. S. Xu, Y. Cheng, K. Gu, Y. Yang, S. Chang, P. Zhou, Jointly attentive spatial-temporal pooling networks for video-based person re-identification, in *Proceedings of the IEEE international conference on computer vision*, (2017), 4733–4742. https://doi.org/10.1109/ICCV.2017.507

27. M. Farenzena, L. Bazzani, A. Perina, V. Murino, M. Cristani, Person re-identification by symmetry-driven accumulation of local features, in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, (2010), 2360–2367. https://doi.org/10.1109/CVPR.2010.5539926

28. D. Gray, H. Tao, Viewpoint invariant pedestrian recognition with an ensemble of localized features, in *European Conference on Computer Vision,* , (2008), 262–275. https://doi.org/10.1007/978-3-540-88682-2_21

29. S. Liao, Y. Hu, X. Zhu, S. Z. Li, Person re-identification by local maximal occurrence representation and metric learning, in *2015 IEEE Conference on Computer Vision and Pattern Recognition*, (2015), 2197–2206. https://doi.org/10.1109/CVPR.2015.7298832

30. B. Ma, Y. Su, F. Jurie, Bicov: a novel image representation for person re-identification and face verification, in *2012 British Machine Vision Conference*, (2012), 1–11. https://doi.org/10.5244/C.26.57

31. M. Geng, Y. Wang, T. Xiang, Y. Tian, Deep transfer learning for person re-identification, preprint, arXiv:1611.05244.

32. W. Song, Y. Wu, J. Zheng, C. Chen, F. Liu, Extended global–local representation learning for video person re-identification, *IEEE Access*, **7** (2019), 122684–122696. https://doi.org/10.1109/ACCESS.2019.2937974

33. Q. Xiao, H. Luo, C. Zhang, Margin sample mining loss: A deep learning based method for person re-identification, preprint, arXiv:1710.00478.

34. J. Meng, W. S. Zheng, J.-H. Lai, L. Wang, Deep graph metric learning for weakly supervised person re-identification, *IEEE Trans. Pattern Anal. Mach. Intell.*, **44** (2022), 6074–6093. https://doi.org/10.1109/TPAMI.2021.3084613

35. T. Matsukawa, E. Suzuki, Person re-identification using cnn features learned from combination of attributes, in *2016 23rd international conference on pattern recognition (ICPR)*, (2016), 2428–2433. https://doi.org/10.1109/ICPR.2016.7900000

36. M. Zheng, S. Karanam, Z. Wu, R. J. Radke, Re-identification with consistent attentive siamese networks, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2019), 5728–5737. https://doi.org/10.1109/CVPR.2019.00588
37. Y. Sun, Q. Xu, Y. Li, C. Zhang, Y. Li, S. Wang, et al., Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification, preprint, arXiv:1904.00537.
38. Y. Fu, Y. Wei, Y. Zhou, H. Shi, G. Huang, X. Wang, et al., Horizontal pyramid matching for person re-identification, in *Proceedings of the AAAI conference on artificial intelligence*, **33** (2019), 8295–8302. https://doi.org/10.1609/aaai.v33i01.33018295
39. Z. Ming, Y. Yang, X. Wei, J. Yan, X. Wang, F. Wang, et al., Global-local dynamic feature alignment network for person re-identification, preprint, arXiv:2109.05759.
40. 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*, **33** (2019), 8618–8625. https://doi.org/10.1609/aaai.v33i01.33018618
41. J. Li, J. Wang, Q. Tian, W. Gao, S. Zhang, Global-local temporal representations for video person re-identification, preprint, arXiv:1908.10049.
42. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *2016 IEEE Conference on Computer Vision and Pattern Recognition*, (2016), 770–778. https://doi.org/10.1109/CVPR.2016.90
43. M. Jaderberg, K. Simonyan, A. Zisserman, K. Kavukcuoglu, Spatial transformer networks, *Adv. Neural Inform. Proc. Syst.*, (2015), 2017–2025.
44. C. Eom, G. Lee, J. Lee, B. Ham, Video-based person re-identification with spatial and temporal memory networks, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2021), 12036–12045. https://doi.org/10.1109/ICCV48922.2021.01182
45. 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 *2018 Proceedings of the IEEE conference on computer vision and pattern recognition*, (2018), 5177–5186. https://doi.org/10.1109/CVPR.2018.00543
46. A. Dehghan, S. Modiri Assari, M. Shah, Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, (2015), 4091–4099. https://doi.org/10.1109/CVPR.2015.7299036
47. Z. Zheng, L. Zheng, Y. Yang, Unlabeled samples generated by GAN improve the person re-identification baseline in vitro, in *2017 IEEE International Conference on Computer Vision*, (2017), 3774–3782. https://doi.org/10.1109/ICCV.2017.405
48. W. Wu, J. Liu, K. Zheng, Q. Sun, Z. J. Zha, Temporal complementarity-guided reinforcement learning for image-to-video person re-identification, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2022), 7319–7328. https://doi.org/10.1109/CVPR52688.2022.00717
49. L. Zheng, Y. Yang, A. G. Hauptmann, Person re-identification: Past, present and future, preprint, arXiv:1610.02984.

50. J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, (2009), 248–255. https://doi.org/10.1109/CVPR.2009.5206848
51. H. Luo, Y. Gu, X. Liao, S. Lai, W. Jiang, Bag of tricks and a strong baseline for deep person re-identification, in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, (2019), 1487–1495, 2019. https://doi.org/10.1109/CVPRW.2019.00190
52. Z. Zhong, L. Zheng, G. Kang, S. Li, Y. Yang, Random erasing data augmentation, in *Proceedings of the AAAI conference on artificial intelligence*, **34** (2020), 13001–13008. https://doi.org/10.1609/aaai.v34i07.7000
53. W. Ruan, C. Liang, Y. Yu, Z. Wang, W. Liu, J. Chen, et al., Correlation discrepancy insight network for video re-identification, *ACM Trans. Multimed. Comput. Commun. Appl.*, **16** (2020), 1–21. https://doi.org/10.1145/3402666
54. L. Cheng, X. Y. Jing, X. Zhu, F. Ma, C. H. Hu, Z. Cai, et al., Scale-fusion framework for improving video-based person re-identification performance, *Neural Comput. Appl.*, **32** (2020), 12841–12858. https://doi.org/10.1007/s00521-020-04730-z
55. Z. Liu, Y. Wang, A. Li, Hierarchical integration of rich features for video-based person re-identification, *IEEE Trans. Circuits Syst. Video Technol.*, **29** (2018), 3646–3659. https://doi.org/10.1109/TCSVT.2018.2883995
56. W. Zhang, X. He, X. Yu, W. Lu, Z. Zha, Q. Tian, A multi-scale spatial-temporal attention model for person re-identification in videos, *IEEE Trans. Image Proc.*, **29** (2020), 3365–3373. https://doi.org/10.1109/TIP.2019.2959653
57. Z. Chen, Z. Zhou, J. Huang, P. Zhang, B. Li, Frame-guided region-aligned representation for video person re-identification, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **34** (2020), 10591–10598. https://doi.org/10.1609/aaai.v34i07.6632
58. Y. Wu, O. E. F. Bourahla, X. Li, F. Wu, Q. Tian, X. Zhou, Adaptive graph representation learning for video person re-identification, *IEEE Trans. Image Proc.*, **29** (2020), 8821–8830. https://doi.org/10.1109/TIP.2020.3001693
59. Y. Yan, J. Qin, J. Chen, L. Liu, F. Zhu, Y. Tai, et al., Learning multi-granular hypergraphs for video-based person re-identification, in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, (2020), 2899–2908. https://doi.org/10.1109/CVPR42600.2020.00297

AIMS Press