Mathematical Biosciences and Engineering

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

MBE, 21(7): 6694–6709.
DOI: 10.3934/mbe.2024293
Received: 08 March 2024
Revised: 16 June 2024
Accepted: 05 July 2024
Published: 23 July 2024

Research article

Video-based person re-identification with complementary local and global features using a graph transformer

Hai Lu, Enbo Luo, Yong Feng and Yifan Wang*

Electric Power Research Institute of Yunnan Power Grid Co., Ltd., Kunming 650217, China

* **Correspondence:** Email: wangyifan@yn.csg.cn; Tel: +8618388421732.

Abstract: In recent years, significant progress has been made in video-based person re-identification (Re-ID). The key challenge in video person Re-ID lies in effectively constructing discriminative and robust person feature representations. Methods based on local regions utilize spatial and temporal attention to extract representative local features. However, prior approaches often overlook the correlations between local regions. To leverage relationships among different local regions, we have proposed a novel video person Re-ID representation learning approach based on a graph transformer, which facilitates contextual interactions between relevant region features. Specifically, we construct a local relation graph to model intrinsic relationships between nodes representing local regions. This graph employs the architecture of a transformer for feature propagation, iteratively refining region features and considering information from adjacent nodes to obtain partial feature representations. To learn compact and discriminative representations, we have further proposed a global feature learning branch based on a vision transformer to capture the relationships between different frames in a sequence. Additionally, we designed a dual-branch interaction network based on multi-head fusion attention to integrate frame-level features extracted by both local and global branches. Finally, the concatenated global and local features, after interaction, are used for testing. We evaluated the proposed method on three datasets, namely iLIDS-VID, MARS, and DukeMTMC-VideoReID. Experimental results demonstrate competitive performance, validating the effectiveness of our proposed approach.

Keywords: video; person re-identification; graph; transformer

1. Introduction

Person re-identification (Re-ID) aims to match individuals across different time and camera views, representing a crucial task in intelligent surveillance [1–3]. Early research primarily focused on image-based person Re-ID, emphasizing the exploration of discriminative information in spatial domains. With the evolution of detection sensors, multimodal information has been introduced into the Re-ID

task, leading to the development of various methods to address differences between modalities, such as network structure [4] and auxiliary features [5]. On another front, some studies leverage multi-frame data and propose different approaches to extract temporal information for video-based person Re-ID [5, 6]. In this scenario, when given an unlabeled query video sequence, the task involves extracting discriminative feature representations to retrieve corresponding individuals from an unlabeled gallery of video sequences. However, how to extract discriminative spatiotemporal aggregation features is the key to improving video-based person Re-ID.

Traditionally, to address this challenge, hierarchical convolutional architectures are often employed to progressively update local patterns. Additionally, some attempts utilize attention-based modules to dynamically infer discriminative information from videos. For instance, Wu et al. [7] embedded prior knowledge about body parts into the network architecture through dense non-local region attention. Despite recent success with convolution-based methods, their inherent limitations in modeling spatiotemporal dependencies and aggregating information pose bottlenecks for accuracy improvement.

In recent years, the transformer [8] has gained attention in computer vision due to its exceptional contextual modeling capabilities. The core idea of this model is to construct long-range relationships between local contents through an attention mechanism. Some hybrid network architectures have been proposed to address context modeling in video-based Re-ID. A widely used paradigm involves using a transformer as a post-processing unit along with convolutional neural networks (CNN) as basic feature extractors. For example, Zhang et al. [9] employed a single transformer to fuse frame-level CNN features. Liu et al. [10] further proposed a multi-stream transformer architecture, with each stream emphasizing a specific dimension of video features. However, in hybrid architectures, the 2D CNN encoder limits long-range spatiotemporal interactions between local contents, hindering the exploration of contextual information. Subsequently, to address this issue, some pure transformer-based methods have been introduced into video-based Re-ID. However, existing frameworks are primarily inspired by video understanding and focus on designing architectures for effective spatiotemporal representation learning. Most algorithms still remain limited in extracting information-rich and person-relevant discriminative information from video segments, which is crucial for large-scale matching tasks.

Moreover, graph neural networks (GNNs) have been widely used in some computer vision tasks, introducing the idea of modeling relationships between graph nodes. Recently, the combination of Re-ID with graph models has also been explored. Cheng et al. [11] formulated structured distance relationships as a Laplacian graph using relations between training samples. Barman et al. [12] mapped the ranking process to a graph theory problem. Shen et al. [13] updated features extracted from images by leveraging similarities between different probe galleries. Chen et al. [14] used multiple graphs in a unified conditional random field (CRF) to simulate relationships between local and global similarities. Yan et al. [15] formulated a person search as a graph matching problem, and resolved it by considering contextual information from probe gallery pairs. To address unsupervised Re-ID, Ye et al. [16] incorporated graph matching into an iterative updating process for robust label estimation.

In person Re-ID tasks, graph-based methods typically construct a graph to represent the relationships between training samples, where the graph nodes correspond to images or videos. In our proposed method, we utilize prior knowledge to learn local relational graphs that model the intrinsic

contextual relationships among regions within image sequences, propagating local information between different area features. To facilitate the propagation of local information, the graph convolutional network (GCN) [17] and graph attention network (GAT) [18] have traditionally been used as feature propagation networks. However, GCN updates node features through a neighbor aggregation mechanism, which means that information for each node is only obtained from its direct neighbors [19]. Although increasing the number of layers can expand the range of information transmission, this also introduces higher computational complexity and the risk of overfitting. GAT introduces an attention mechanism to assign different weights to each neighbor, enhancing the model's flexibility in learning the importance of neighbors, but this also results in higher computational demands, particularly in graphs with high node degrees [20]. The transformer [8], through its self-attention mechanism, can directly compute the relationship between any two positions in a sequence, capturing global information and solving long-distance dependency issues. This capability is crucial for learning the contextual relationships of graph nodes and extracting person discriminative information from video segments. Therefore, in the proposed method, we leverage transformers to update local relational graphs, thereby modeling the intrinsic contextual relationships between regions in image sequences. The main contributions of this work can be summarized as follows.

- We introduce a video person Re-ID framework based on a graph transformer, which effectively integrates both local and global features. This framework is designed to learn feature representations that capture rich semantics and discriminative information essential for person identification.
- We present a local feature learning approach based on a graph transformer, facilitating contextual interactions among relevant region features. Concretely, we construct a local relation graph using local regions to model intrinsic relationships between graph nodes. The transformer architecture is used on the local relation graph for feature propagation, iteratively refining regional features while considering information from adjacent nodes for feature representation.
- To learn compact and discriminative representations, we further propose a global feature learning branch based on a vision transformer to capture the relationships between different frames in a sequence. Additionally, we design a dual-branch interaction network based on multi-head fusion attention to integrate frame-level features extracted by local and global branches. This enhances the distinctiveness and richness of semantic information in video sequence-level features. Finally, the global features and the updated local region features are concatenated for testing.
- We conducted extensive experiments on three widely adopted benchmarks, namely iLIDS-VID, MARS, and DukeMTMCVideoReID. The experimental results affirm the effectiveness of our proposed approach.

2. Related works

2.1. Image-based person re-identification

Image-based person Re-ID primarily focuses on acquiring effective representations of individuals [21–23]. Early approaches were dominated by meticulously designed manual feature extraction. Recently, the field has witnessed a surge in the adoption of deep learning as the mainstream methodology for representation learning in person Re-ID. CNNs have emerged as

prevalent feature extractors in this context. Notably, OSNet [24] integrates multi-scale features into an attentional subnetwork, producing information-rich full-scale features. Other studies [25, 26] place emphasis on extracting and aligning semantic information to tackle alignment issues arising from pose/viewpoint variations and imperfect person detection. To counter the detrimental impact of noisy labels, Ye et al. [27] introduced a self-label refining strategy that intricately combined label optimization with deep network training. Additionally, several image-based person Re-ID methods have explored using the vision transformer (ViT) [8] to improve the performance. For instance, TransReID [28] employs the ViT as the backbone to extract discriminative features from randomly sampled patch groups.

2.2. Video-based person re-identification

Compared to image-based person Re-ID, video-based person Re-ID often exhibits superior performance due to the inclusion of temporal information and the utilization of multiple frames to alleviate occlusion. Traditional video-based Re-ID methods typically focus on two aspects to obtain more robust and distinctive representations from frame sequences: (1) encoding temporal information and (2) aggregating temporal information.

To encode additional temporal information, reference [6] directly used time information as additional features. Some approaches employed recurrent models, such as RNN [29] and LSTM [30], to handle temporal information. Further advancements, such as [31], introduced attention mechanisms to dynamically fuse temporal features. Another category of methods introduced optical flow to capture temporal motion [32]. Additionally, spatiotemporal pooling was performed directly on video sequences, generating global representations through CNNs [33]. Recently, 3D CNNs have been employed to encode video features in a joint spatiotemporal manner [34]. M3D [35] endowed 2D CNNs with multi-scale temporal feature extraction capabilities through multi-scale three-dimensional convolutional kernels.

To generate discriminative features from complete video features, literature [36] used average pooling along the temporal dimension to aggregate spatiotemporal feature maps. Recently, an attention-based approach dynamically highlighted different video frames or regions, filtering out more discriminative features from these key frames or regions and significantly improving performance. For example, Liu et al. [10] introduced cross attention to aggregate multi-view video features through pairwise interactions between these views. In addition to exploring more effective architectural designs, Zhao et al. [37] investigated pedestrian attributes, such as shoes, bags, down jacket color, or gait (the walking style of pedestrians), to provide a more comprehensive description of pedestrian features. Chang et al. [38] tightly integrated gait recognition and video-based Re-ID as coherent tasks using a hybrid framework that includes a set-based gait recognition branch. Some studies embedded attribute predictors into the network, supported by annotations obtained from pretraining on attribute datasets. For instance, Chai et al. [39] categorized attributes into ID-related and ID-unrelated attributes, proposing a new triple loss method for pose and motion invariant learning to mine the most challenging samples considering pose and motion status distances.

Although the mentioned methods have achieved significant advancements in performance, the transformer is considered a more powerful architecture for sequence data processing, which may elevate the performance ceiling of video-based Re-ID. To illustrate this, the transformer can easily adapt to video data, supporting global attention mechanisms to capture spatiotemporal dependencies

and time-position encoding to sequence spatiotemporal positions. Additionally, class tokens are readily available for transformer-based models to aggregate spatiotemporal information. However, transformers have several drawbacks [9], and currently, there are relatively few works on transformer-based video-based person Re-ID. In this work, we aim to explore the potential of the graph and transformer in video-based person Re-ID.

3. Method

The purpose of video-based person Re-ID is to retrieve gallery video sequences with the same identity as a given query video sequence. The overall architecture of our proposed method is illustrated in Figure 1. For a specific identity's video sequence, we employ a constrained sampling method [40] to randomly sample T frames, which are then grouped into an image sequence $\{I_t\}_{t=1,\ldots,T}$. Initially, these sequences are input into a feature extraction module based on ResNet50, where the stride of the first residual block in Conv-5 is set to 1. In the global branch, generalized mean pooling is applied to the feature map to generate a video representation denoted by $x \in \mathbb{R}^{b\times T\times d}$, where b represents the batch size, T represents the number of frames, and d represents the feature dimension. In the local branch, we employ pyramid pooling to obtain regional features $X = \{x_i\}_{i=1}^{T\times N}$, where the feature map is vertically divided into 1, 2, and 4 regions in our experiments, with $N = 7$ representing the number of regions for a single frame. Subsequently, we construct a graph of local region relationships using feature similarity, capturing intrinsic relationships among region features. In the graph feature propagation module, region features are iteratively updated by aggregating contextual information from neighboring regions on the graph. Following this, we utilize average pooling to generate a video representation. The network is supervised by a combination of cross-entropy loss and triplet loss.

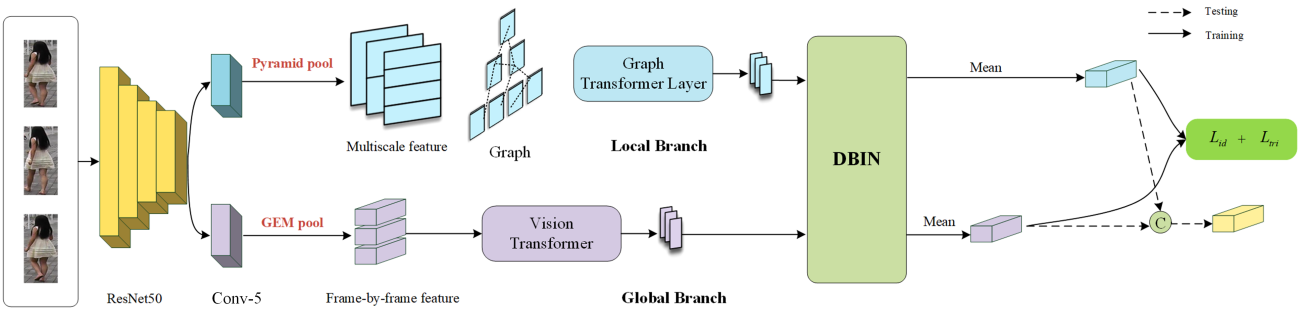

Figure 1. The overall architecture of our approach comprises a local branch and a global branch. The local branch primarily focuses on modeling intrinsic relationships among local regions to extract fine-grained information. The global branch is designed to capture relationships between different frames in the sequence, facilitating the learning of a global representation for the entire video sequence. DBIN is used to fuse frame-level features extracted from local and global branches.

3.1. Local feature network

The relationships between different parts of the human body are beneficial for mitigating the impact of complexities such as occlusion and background noise. Therefore, describing the relationships between different body parts and propagating contextual information are crucial for learning discriminative sequence features. Graphs are commonly used to model such relationships, and we adopt GNNs to mine information between regions.

3.1.1. Local relation graph

To describe the relationships between regions, as illustrated in Figure 1, we propose to learn a local relation graph $G = \{V, A\}, V = \{v_i\}_{i=1}^{T\times N}$, which is a set of features containing $T \times N$ nodes aimed at capturing the affinity between regions. Each node v_i corresponds to a spatial region in a frame. For two nodes v_i and v_j, the node features are denoted as x_i and x_j. $S(x_i, x_j)$ is used to calculate the cosine similarity between nodes. The adjacency matrix A is represented as:

$$A_{ij} = S(x_i, x_j) = \frac{2}{e^{\|x_i - x_j\|_2} + 1} \tag{3.1}$$

Next, we utilize pre-defined graph structures to generate position encoding for graph nodes:

$$\Delta = 1 - D^{-1/2} A D^{-1/2} = U^T \Lambda U \tag{3.2}$$

where A and D represent the adjacency matrix and degree matrix of the local relation graph, respectively. Λ and U denote the Laplacian eigenvalue matrix and eigenvector matrix. $U^T \Lambda U$ is the factorization of the graph Laplacian matrix, and the local relation graphs in the same dataset share the same initialized adjacency matrix. Following [41], we utilize the K smallest singular eigenvectors as the node position encoding, denoted as $e_i \in \mathbb{R}^K$. v_i and e_i are mapped to a feature space of the same dimensionality d through an affine transformation and summed up, as defined:

$$v_i^l = (W_v v_i + b_v) + (W_p e_i + b_p) \tag{3.3}$$

where $v_i^l \in \mathbb{R}^d$ represents the i-th positional encoding node, and $W_v \in \mathbb{R}^{d\times N}$, $W_p \in \mathbb{R}^{d\times K}$, b_v, $b_p \in \mathbb{R}^d$ are the learnable parameters of the linear mapping layer for the i-th node v_i and its corresponding positional encoding. It is important to note that Laplacian position encoding is added only to the node features of the input layer and not to the intermediate graph transformer layers.

3.1.2. Transformer-based graph feature propagation network

After obtaining the graph, contextual information is propagated, iteratively updating the original spatial region features. As shown in Figure 2, we employ a transformer architecture adapted for graph input to aggregate information from adjacent nodes for each node. In the graph feature propagation network, we use a layer-wise transformer architecture, with the number of heads for multi-head attention being K. At the k-th attention head of the l-th layer, the feature aggregation and updating operations are defined as follows:

$$\hat{v}_i^l = O^l \|_{k=1}^K \left(\sum_{j\in\mathcal{N}_i} w_{ij}^{k,l} V^{k,l} v_j^l \right) \tag{3.4}$$

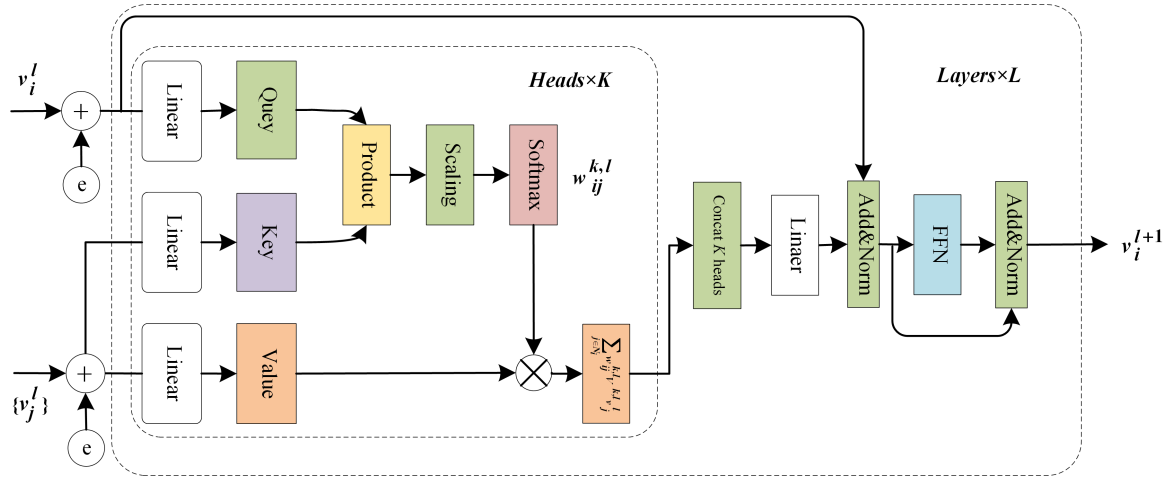

Figure 2. Diagram of the graph transformer layer using Laplacian eigenvectors. e denotes positional encoding, which is added to the input node embeddings before features are fed into the first layer.

In the equation, $w_{ij}^{k,l} = \text{softmax}\left(\frac{Q^{k,l}v_i^l K^{k,l}v_j^l}{\sqrt{d_k}}\right)$. $Q^{k,l}$, $K^{k,l}$, $V^{k,l} \in \mathbb{R}^{d_k \times d}$, $O^l \in \mathbb{R}^{d \times d}$ represent the learnable parameters of the linear mapping layer. $k = 1, \ldots, K$ denotes the number of attention heads, and $\|$ represents the concatenation operation. For clarity, we use $\hat{v}_i^l \in \mathbb{R}^d$ to represent the i-th node, which denotes the node features learned from different heads connected starting from the first layer. It is noteworthy that the graph transformer layer naturally extends self-attention [31] to learn relationships among graph nodes and can be seen as a general paradigm for simultaneously capturing relationships between adjacent and non-adjacent elemental nodes in the graph representing local region relationships. This allows the integration of more crucial features from relevant nodes into the final node representation. Finally, following [41], we apply a feed-forward network (FFN) with residual connections and batch normalization:

$$\begin{aligned} \overline{v}_i^l &= \text{Norm}\left(v_i^l + \hat{v}_i^l\right), \\ v_i^{l+1} &= \text{Norm}\left(\overline{v}_i^l + W_2^l \sigma\left(W_1^l \overline{v}_i^l\right)\right), \end{aligned} \tag{3.5}$$

where Norm$(\cdot)$ represents the batch normalization operation. $W_1^l \in \mathbb{R}^{2d \times d}$, $W_2^l \in \mathbb{R}^{d \times 2d}$ are the learnable parameter of the FFN layer. $\sigma(\cdot)$ is the ReLU activation function. $\overline{v}_i^l$ and v_i^{l+1} respectively represent the intermediate node features and the final output node features of the l-th graph transformer layer. We obtain the graph representation corresponding to each node by taking the average of the node features in each local relation graph. These continuous graph representations are then integrated into the final sequence-level graph representation S, where N represents the number of local relation graphs:

$$S = \frac{1}{T}\sum_{t=1}^{T} s^t = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{N}\sum_{i=1}^{N} v_i^t \tag{3.6}$$

where $S, s^t \in \mathbb{R}^d$ represent the video sequence features and features of a single frame image, respectively.

3.2. Global feature network

Transformer is a powerful neural network architecture initially designed for natural language processing tasks but has been successfully applied in the field of computer vision. In the context of video-based person Re-ID, using the transformer enables effective feature learning and representation for individuals across video sequences. In this paper, we first feed the output of the last residual block of ResNet50 through generalized mean pooling to vectorize the feature maps, obtaining a feature vector for each frame. Let $F = \{f_i\}_{i=1}^{T}$ be the set of feature vectors, and $f_i \in \mathbb{R}^{b\times T\times d}$ represents the i-th frame in the video sequence. Subsequently, each extracted frame-level feature is treated as a token vector and fed into the ViT to model the entire video sequence. The self-attention mechanism of the transformer is utilized to capture relationships between different frames in the sequence. Consequently, each token vector, after the vision transformer interacts with the remaining token vectors, captures information from other frames. Finally, averaging all output token vectors yields the global feature for the entire video sequence. This can be formally defined as follows:

$$F = \frac{1}{T}\sum_{t=1}^{T} f^t \tag{3.7}$$

where $F, f^t \in \mathbb{R}^d$ represent the video sequence features and frame-level features, respectively.

3.3. The dual-branch interaction network with multi-head fused attention

For the purpose of incorporating richer semantic information into the sequence-level features used for final testing, we propose a dual-branch interaction network (DBIN) based on multi-head fused attention. DBIN is designed to fuse frame-level features extracted from local and global branches. The structure is illustrated in Figure 3. Assuming F_l and F_g represent the feature sets of each frame from the local and global branches, the feature sequences of the two branches are taken as inputs to this interaction network. We perform dimension reduction on the input features to obtain two sequences with dimensions $T \times D$. Subsequently, layer normalization is applied to the features. The overall process is described as follows:

$$\begin{cases} y_1 = LN(LP(F_l)) \\ y_2 = LN(LP(F_g)) \end{cases} \tag{3.8}$$

where LP represents the linear mapping operation, LN denotes the layer normalization, and $y_1, y_2 \in \mathbb{R}^{d\times N}$ represents the two sequences of features after dimension reduction and layer normalization.

The two output feature maps of the DBIN can be formulated as:

$$\begin{cases} z_1 = \text{Attention}\,(Q_1, K_1, V_2) = \text{softmax}\left(\frac{Q_1K_1}{\sqrt{d_k}}\right)V_2 \\ z_2 = \text{Attention}\,(Q_2, K_2, V_1) = \text{softmax}\left(\frac{Q_2K_2}{\sqrt{d_k}}\right)V_1 \end{cases} \tag{3.9}$$

where Q represents the query, K represents the key, V is the value, and d_k represents the dimensionality of the input data. DBIN differs from the traditional self-attention mechanism. The goal of the DBIN is to capture the correlation between the query Q and the key K, obtain an attention map, and then derive the feature values based on this attention map. This cross-attention mechanism is more advantageous for the information interaction between the two feature maps in fusion tasks, thereby improving the fusion effectiveness.

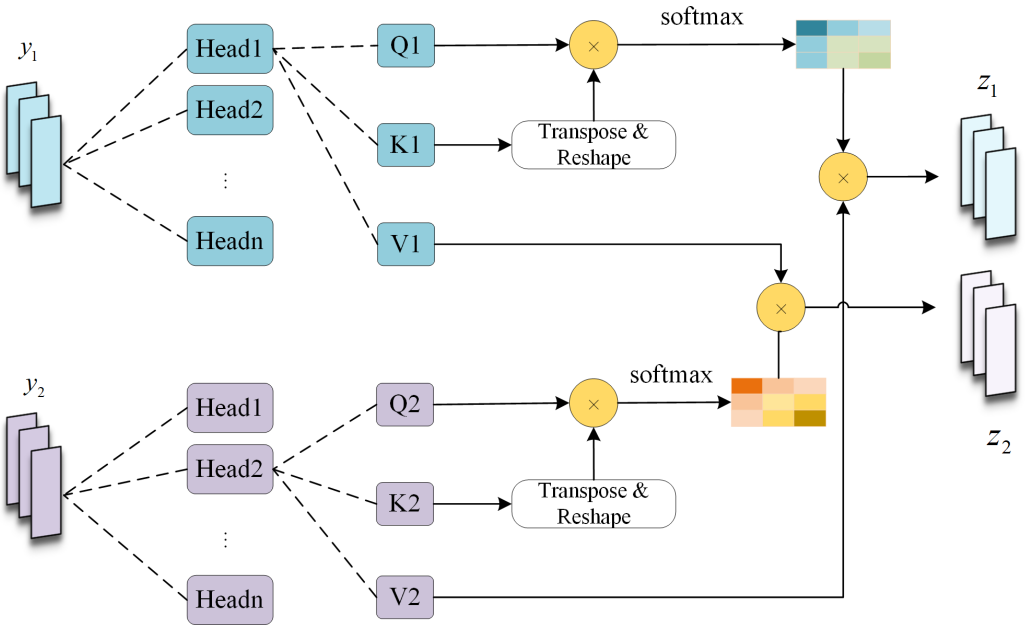

Figure 3. A dual-branch interaction network based on multi-head fusion attention (DBIN), used to integrate frame-level features extracted by the local branch and the global branch, so that the final extracted features contain rich semantic information.

4. Experiments and results

4.1. Dataset and evaluation scheme

In this paper, we evaluate our proposed method MSTAT on three widely used video-based person Re-ID datasets: iLIDS-VID [42], DukeMTMC-VideoReID (DukeV) [43], and MARS [44]. iLIDS-VID consists of 600 video sequences of 300 individuals captured by two cameras. The number of frames in these video sequences varies from 23 to 192. The test set shares 150 identities with the training set. DukeMTMC-VideoReID is a large-scale video-based dataset with 4832 videos from 1404 identities. In the subsequent sections, we use the abbreviation "DukeV" for the DukeMTMC-VideoReID dataset. Video sequences in the DukeV dataset are generally longer than those in other datasets, with an average of 168 frames per sequence. MARS is one of the largest video Re-ID datasets, collecting 20,000 video sequences from 1261 identities captured by six cameras. Frames in the video sequences are relatively less aligned in the tracker as they are obtained from deformable part model (DPM) detectors and generalized maximum multi clique problem (GMMCP) trackers rather than manually annotated [44]. Additionally, the dataset includes around 3200 distractor sequences to simulate real-world scenarios.

To evaluate the MARS and DukeV datasets, we used two metrics: the cumulative match characteristic (CMC) curve and the mean average precision (mAP). However, for the gallery set of iLIDS-VID, only the cumulative accuracy is provided for this benchmark.

4.2. Experimental setup

All experiments were conducted on a single RTX 3090 GPU. Optimization was performed using the stochastic gradient descent (SGD) optimizer. The learning rate was initialized at 0.1 and linearly adjusted through a warm-up strategy in the first 10 epochs. It was then reduced to 0.01 at the 35th epoch and to 0.001 at the 80th epoch. The weight decay was set to 0.0005, and the batch size was 8 (each mini-batch included 4 identities, with each identity having 2 video sequences). Training was performed for over 200 epochs. During the training phase, a constrained random sampling strategy was employed, randomly extracting $T = 8$ frames from each video and grouping them into video sequences.

4.3. Comparison with existing methods

To validate the effectiveness of our proposed method, we compare it with several state-of-the-art methods on iLIDS-VID, MARS, and Duke-V, including STMP [45], SCAN [46], AP3D [34], TCLNet [47], GRL [48], STRF [49], STT [9], Snippet+OF [50], SCAN+OF [46], DCCT [51], MSTAT [52], and PiT [53]. The experimental results on MARS and iLIDS-VID are listed in Table 1 and those on Duke-V are shown in Table 2. On the MARS dataset, our method outperforms the previous best method, DCCT, by 0.3% in Rank-1. The proposed method achieves a good performance on the Duke-V dataset, with accuracies of 97.5% for Rank-1 and 97.8% for mAP. The mAP metric surpasses that of the state-of-the-art DCCT. The experimental results confirm the effectiveness and superiority of our proposed method.

Table 1. Comparison with state-of-the-art methods on MARS and iLIDS-VID datasets, providing Rank-1, -5, -20 accuracies (%) and mAP (%). The experimental results demonstrate that the proposed method achieves state-of-the-art performance.

Method	Source	MARS				iLIDS-VID		
		R-1	R-5	R-20	mAP	R-1	R-5	R-20
STMP [45]	AAAI2019	84.4	93.2	96.3	72.7	84.3	**96.8**	99.5
SCAN [46]	TIP19	86.6	94.8	97.1	76.7	81.3	93.3	98.0
AP3D [34]	ECCV20	90.7	-	-	85.6	88.7	-	-
TCLNet [47]	ECCV20	89.8	-	-	85.1	86.6	-	-
GRL [48]	CVPR21	90.4	96.7	-	84.8	90.4	98.3	-
STRF [49]	ICCV21	90.3	-	-	86.1	89.3	-	-
STT [9]	Arxiv21	88.7	-	-	86.3	87.5	95.0	-
PiT [53]	TII22	90.2	97.2	-	86.8	92.0	**98.9**	100.0
DCCT [51]	TNNLS23	92.3	-	-	**87.5**	91.7	98.6	-
MSTAT [52]	TMM23	91.8	97.4	-	86.5	93.3	99.3	-
Snippet+OF [50]	CVPR2018	86.3	94.7	98.2	76.1	85.4	96.7	-
SCAN+OF [46]	TIP19	87.2	95.2	98.1	77.2	88	96.7	100.0
Graph Trans (ours)	-	**92.5**	**97.5**	**98.5**	86.4	**93.8**	98.6	**100.0**

Results on iLIDS-VID, as shown in Table 1, indicate the superiority of the proposed method over existing state-of-the-art methods on this dataset. Specifically, our method achieves a Rank-1 accuracy of 93.8%, surpassing all previous methods, even without considering optical flow. On this small-scale

dataset, the comparison between Snippet and Snippet+OF demonstrates that motion information provides more reliable features than appearance cues. Even when compared to methods utilizing optical flow, our proposed method remains competitive.

Table 2. Comparison with state-of-the-art methods on Duke-V datasets, providing Rank-1, -5, -20 accuracies (%) and mAP (%). The experimental results demonstrate that the proposed method achieves state-of-the-art performance.

Method	Source	R-1	R-5	R-20	mAP
STMP [45]	AAAI2019	-	-	-	-
AP3D [34]	ECCV20	97.2	-	-	96.1
TCLNet [47]	ECCV20	96.9	-	-	96.2
GRL [48]	CVPR21	95.0	98.7	-	93.8
STRF [49]	ICCV21	97.4	-	-	96.4
STT [9]	Arxiv21	97.6	-	-	97.4
DCCT [51]	TNNLS23	**98.4**	-	-	97.6
MSTAT [52]	TMM23	97.4	**99.3**	-	96.4
Graph Trans (ours)	-	97.5	98.7	**99.3**	**97.8**

4.4. Ablation experiment

To analyze the effectiveness of each component in the proposed method, we conduct the ablation experiment on the MARS dataset. The experimental results are presented in Table 3. In the ablation experiment, the baseline includes only the ResNet backbone and 3D global average pooling, supervised by a cross-entropy loss and a triplet loss. The Rank-1 and mAP accuracies of the baseline method are 87.9% and 77.7%, respectively. The baseline+local branch indicates the adoption of the graph transformer module in the local relation graph branch, achieving Rank-1 and mAP accuracies of 89.6% and 82.9%, respectively. The baseline+global branch refers to the additional branch of the framework for video sequence-level feature learning based on the vision transformer.

Table 3. Ablation experiment results on the MARS dataset.

Method	R-1	R-5	R-20	mAP
Baseline (Resnet50)	87.9	95.8	97.4	77.7
Baseline+local branch	89.6	96.5	97.4	82.9
Baseline+global branch	89.0	95.8	97.0	80.9
Baseline+local branch+local branch	90.0	96.5	98.5	84.4
Baseline+local branch+local branch+DBIN	**92.5**	**97.5**	**98.5**	**86.4**

Compared to using the global branch alone, the addition of the local branch improves the accuracy of Rank-1 and mAP by 0.6% and 2.0%, respectively. Moreover, by combining both the global and local branches, we achieve 90.0% and 84.4% on the MARS dataset. After integrating the DBIN module, there is an increase of 2.5% and 2.0% in the accuracy of Rank-1 and mAP, respectively. Utilizing these components, we elevate the accuracy of Rank-1 and mAP from 87.9% and 77.7% to 92.5% and 86.4%, respectively.

5. Conclusions

This paper proposes a novel video-based person Re-ID representation learning approach based on graph transformer. The proposed method learns local region relation graphs in spatial regions. By aggregating contextual information from neighboring nodes, it captures intrinsic relational structural information among person feature nodes. Furthermore, it utilizes the transformer architecture to propagate complementary contextual information, enriching person feature representations. Additionally, we further propose a global feature learning branch based on ViT to capture the relationships between different frames in a sequence. A dual-branch interaction network, designed on the principle of multi-head fusion attention, integrates frame-level features extracted by both local and global branches. Ultimately, the concatenated global and local features, post-interaction, are utilized for testing, which is conducive to learning compact and discriminative representations. Experimental results on three public datasets demonstrate the effectiveness of this approach, and the ablation study investigates the contribution of the proposed components. Additionally, our current framework only constructs local relation graphs that consider contextual relationships within single images, but it does not fully explore the relationships between frames in video sequences. In future efforts, we plan to more comprehensively utilize both intra-frame and inter-frame relationships to develop a more robust graph structure, enhancing the accuracy of person identification.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. H. Li, K. Xu, J. Li, Z. Yu, Dual-stream reciprocal disentanglement learning for domain adaptation person re-identification, *Knowl. Based Syst.*, **251** (2022), 109315. https://doi.org/10.1016/j.knosys.2022.109315
2. S. Yan, Y. Zhang, M. Xie, D. Zhang, Z. Yu, Cross-domain person re-identification with pose-invariant feature decomposition and hypergraph structure alignment, *Neurocomputing*, **467** (2022), 229–241. https://doi.org/10.1016/j.neucom.2021.09.054
3. H. Li, J. Xu, Z. Yu, J. Luo, Jointly learning commonality and specificity dictionaries for person re-identification, *IEEE Trans. Image Process.*, **29** (2020), 7345–7358. https://doi.org/10.1109/TIP.2020.3001424
4. Y. Zhang, Y. Wang, H. Li, S. Li, Cross-compatible embedding and semantic consistent feature construction for sketch re-identification, in *Proceedings of the 30th ACM International Conference on Multimedia (MM'22)*, (2022), 3347–3355. https://doi.org/10.1145/3503161.3548224

5. H. Li, M. Liu, Z. Hu, F. Nie, Z. Yu, Intermediary-guided bidirectional spatial-temporal aggregation network for video-based visible-infrared person reidentification, *IEEE Trans. Circuits Syst. Video Technol.*, **33** (2023), 4962–4972. https://doi.org/10.1109/TCSVT.2023.3246091
6. A. Subramaniam, A. Nambiar, A. Mittal, Co-segmentation inspired attention networks for video-based person re-identification, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, IEEE, (2019), 562–572. https://doi.org/10.1109/ICCV.2019.00065
7. D. Wu, M. Ye, G. Lin, X. Gao, J. Shen, Person re-identification by context-aware part attention and multi-head collaborative learning, *IEEE Trans. Inf. Forensics Secur.*, **17** (2021), 115–126. https://doi.org/10.1109/TIFS.2021.3075894
8. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, et al., An image is worth 16x16 words: Transformers for image recognition at scale, in *9th International Conference on Learning Representations (ICLR)*, (2021). https://doi.org/10.48550/arXiv.2010.11929
9. T. Zhang, L. Wei, L. Xie, Z. Zhuang, Y. Zhang, B. Li, et al., Spatiotemporal transformer for video-based person re-identification, preprint, arXiv:2103.16469.
10. X. Liu, P. Zhang, C. Yu, H. Lu, X. Qian, X. Yang, A video is worth three views: Trigeminal transformers for video-based person re-identification, preprint, arXiv:2104.01745.
11. D. Cheng, Y. Gong, X. Chang, W. Shi, A. Hauptmann, N. Zheng, Deep feature learning via structured graph laplacian embedding for person re-identification, *Pattern Recognit.*, **82** (2018), 94–104. https://doi.org/10.1016/j.patcog.2018.05.007
12. A. Barman, S. K. Shah, Shape: A novel graph theoretic algorithm for making consensus-based decisions in person re-identification systems, in *Proceedings of the IEEE International Conference on Computer Vision*, IEEE, (2017), 1115–1124. https://doi.org/10.1109/ICCV.2017.127
13. Y. Shen, H. Li, S. Yi, D. Chen, X. Wang, Person re-identification with deep similarity-guided graph neural network, in *Proceedings of the European Conference on Computer Vision (ECCV)*, Springer, (2018), 486–504. https://doi.org/10.48550/arXiv.1807.099757
14. D. Chen, D. Xu, H. Li, N. Sebe, X. Wang, Group consistent similarity learning via deep crf for person re-identification, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, (2018), 8649–8658. https://doi.org/10.1109/CVPR.2018.00902
15. Y. Yan, Q. Zhang, B. Ni, W. Zhang, M. Xu, X. Yang, Learning context graph for person search, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, (2019), 2153–2162. https://doi.org/10.1109/CVPR.2019.00226
16. M. Ye, A. J. Ma, L. Zheng, J. Li, P. C. Yuen, Dynamic label graph matching for unsupervised video re-identification, in *Proceedings of the IEEE International Conference on Computer Vision*, IEEE, (2017), 5122–5160. https://doi.org/10.1109/ICCV.2017.550
17. L. Bao, B. Ma, H. Chang, X. Chen, Preserving structural relationships for person re-identification, in *2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, IEEE, (2019), 120–125. https://doi.org/10.1109/ICMEW.2019.00028

18. Z. Zhang, H. Zhang, S. Liu, Person re-identification using heterogeneous local graph attention networks, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, (2021), 12131–12140. https://doi.org/10.1109/CVPR46437.2021.01196

19. T. N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, in *International Conference on Learning Representations (ICLR)*, (2016). https://doi.org/10.48550/arXiv.1609.02907

20. P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio, Graph attention networks, preprint, arXiv:1710.10903.

21. H. Li, S. Yan, Z. Yu, D. Tao, Attribute-identity embedding and self-supervised learning for scalable person re-identification, *IEEE Transactions on Circuits and Systems for Video Technology*, **30** (2020), 3472–3485. https://doi.org/10.1109/TCSVT.2019.2952550

22. 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** (2022), 2814–2830. https://doi.org/10.1109/TCSVT.2021.3099943

23. 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. Forensics Secur.*, **16** (2021), 1480–1494. https://doi.org/10.1109/TIFS.2020.3036800

24. K. Zhou, Y. Yang, A. Cavallaro, T. Xiang, Omni-scale feature learning for person re-identification, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, IEEE, (2019), 3701–3711. https://doi.org/10.1109/ICCV.2019.00380

25. F. Yu, X. Jiang, Y. Gong, S. Zhao, X. Guo, W. S. Zheng, et al., Devil's in the details: Aligning visual clues for conditional embedding in person re-identification, preprint, arXiv:2009.05250.

26. Z. Zhang, C. Lan, W. Zeng, Z. Chen, Densely semantically aligned person re-identification, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, (2019), 667–676. https://doi.org/10.1109/CVPR.2019.00076

27. M. Ye, H. Li, B. Du, J. Shen, L. Shao, S. C. H. Hoi, Collaborative refining for person re-identification with label noise, *IEEE Trans. Image Process.*, **31** (2021), 379–391. https://doi.org/10.1109/TIP.2021.3131937

28. S. He, H. Luo, P. Wang, F. Wang, H. Li, W. Jiang, Transreid: Transformer-based object re-identification, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, IEEE, (2021), 14993–15002. https://doi.org/10.1109/ICCV48922.2021.01474

29. T. Mikolov, M. Karafiàt, L. Burget, J. Cernockỳ, S. Khudanpur, Recurrent neural network based language model, in *11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010*, (2010), 1045–1048. https://doi.org/10.21437/Interspeech.2010-343

30. S. Hochreiter, J. Schmidhuber, Long short-term memory, *Neural Comput.*, **9** (1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

31. N. McLaughlin, J. Martinez 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*, IEEE, (2016), 1325–1334. https://doi.org/10.1109/CVPR.2016.148

32. D. Chung, K. Tahboub, E. J. Delp, A two stream siamese convolutional neural network for person re-identification, in *Proceedings of the IEEE International Conference on Computer Vision*, IEEE, (2017), 1992–2000. https://doi.org/10.1109/ICCV.2017.218

33. Y. Suh, J. Wang, S. Tang, T. Mei, K. M. Lee, Part-aligned bilinear representations for person re-identification, in *Proceedings of the European Conference on Computer Vision (ECCV)*, Springer, (2018), 418–437. https://doi.org/10.1007/978-3-030-01264-9_25

34. X. Gu, H. Chang, B. Ma, H. Zhang, X. Chen, Appearance-preserving 3D convolution for video-based person re-identification, in *Proceedings of the European Conference on Computer Vision (ECCV)*, Springer, (2020), 228–243. https://doi.org/10.1007/978-3-030-58536-5_14

35. 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*, (2019), 8618–8625. https://doi.org/10.1609/aaai.v33i01.33018618

36. P. Zhang, J. Xu, Q. Wu, Y. Huang, X. Ben, Learning spatial-temporal representations over walking tracklet for long-term person re-identification in the wild, *IEEE Trans. Multimedia*, **23** (2020), 3562–3576. https://doi.org/10.1109/TMM.2020.3028461

37. 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*, IEEE, (2019), 4908–4917. https://doi.org/10.1109/CVPR.2019.00505

38. Z. Chang, Z. Yang, Y. Chen, Q. Zhou, S. Zheng, Seq-masks: Bridging the gap between appearance and gait modeling for video-based person re-identification, in *2021 International Conference on Visual Communications and Image Processing (VCIP)*, IEEE, (2021), 1–5. https://doi.org/10.1109/VCIP53242.2021.9675368

39. T. Chai, Z. Chen, A. Li, J. Chen, X. Mei, Y. Wang, Video person re-identification using attribute-enhanced features, *IEEE Trans. Circuits Syst. Video Technol.*, **32** (2022), 7951–7966. https://doi.org/10.1109/TCSVT.2022.3189027

40. L. Wu, Y. Wang, L. Shao, M. Wang, 3-d personvlad: Learning deep global representations for video-based person reidentification, *IEEE Trans. Neural Networks Learn. Syst.*, **30** (2019), 3347–3359. https://doi.org/10.1109/TNNLS.2019.2891244

41. V. Dwivedi, X. Bresson, A generalization of transformer networks to graphs, preprint, arXiv:2012.0969.

42. T. Wang, S. Gong, X. Zhu, S. Wang, Person re-identification by video ranking, in *Proceedings of the European Conference on Computer Vision (ECCV)*, Springer, (2014), 688–703. https://doi.org/10.1007/978-3-319-10593-2_45

43. E. Ristani, F. Solera, R. Zou, R. Cucchiara, C. Tomasi, Performance measures and a data set for multi-target, multi-camera tracking, in *Proceedings of the European Conference on Computer Vision (ECCV)*, Springer, (2016), 17–35. https://doi.org/10.1007/978-3-319-48881-3_2

44. 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, Springer, (2016), 868–884. https://doi.org/10.1007/978-3-319-46466-4_52
45. Y. Liu, Z. Yuan, W. Zhou, H. Li, Spatial and temporal mutual promotion for video-based person re-identification, in *Proceedings of the AAAI Conference on Artificial Intelligence*, (2019), 8786–8793. https://doi.org/10.1609/aaai.v33i01.33018786
46. R. Zhang, J. Li, H. Sun, Y. Ge, P. Luo, X. Wang, et al., Scan: Self-and-collaborative attention network for video person re-identification, *IEEE Trans. Image Process.*, **28** (2019), 4870–4882. https://doi.org/10.1109/TIP.2019.2911488
47. G. Chen, Y. Rao, J. Lu, J. Zhou, Temporal coherence or temporal motion: Which is more critical for video-based person re-identification?, in *Proceedings of the European Conference on Computer Vision (ECCV)*, Springer, (2020), 660–676. https://doi.org/10.1007/978-3-030-58598-3_39
48. X. Liu, P. Zhang, C. Yu, H. Lu, X. Yang, Watching you: Global-guided reciprocal learning for video-based person re-identification, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, (2021), 13329–13338. https://doi.org/10.1109/CVPR46437.2021.01313
49. A. Aich, M. Zheng, S. Karanam, T. Chen, A. K. Roy-Chowdhury, Z. Wu, Spatio-temporal representation factorization for video-based person re-identification, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, IEEE, (2021), 152–162. https://doi.org/10.1109/ICCV48922.2021.00022
50. 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*, IEEE, (2018), 1169–1178. https://doi.org/10.1109/CVPR.2018.00128
51. X. Liu, C. Yu, P. Zhang, H. Lu, Deeply coupled convolution–transformer with spatial–temporal complementary learning for video-based person re-identification, *IEEE Trans. Neural Networks Learn.Syst.*, (2023), 1–11. https://doi.org/10.1109/TNNLS.2023.3271353
52. Z. Tang, R. Zhang, Z. Peng, J. Chen, L. Lin, Multi-stage spatio-temporal aggregation transformer for video person re-identification, *IEEE Trans. Multimedia*, **25** (2023), 7917–7929. https://doi.org/10.1109/TMM.2022.3231103
53. X. Zang, G. Li, W. Gao, Multidirection and multiscale pyramid in transformer for video-based pedestrian retrieval, *IEEE Trans. Ind. Inf.*, **18** (2022), 8776–8785. https://doi.org/10.1109/TII.2022.3151766