A Sparse Graph Wavelet Convolution Neural Network for Video-based Person Re-identification

Yingmao Yao[a], Xiaoyan Jiang[a,*], Hamido Fujita[b,c,d], Zhijun Fang[a]

[a]*Shanghai University of Engineering Science, Shanghai, China*
[b]*i-SOMET Incorporated Association, Morioka, Japan*
[c]*Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI), University of Granada, Granada, Spain*
[d]*Iwate Prefectural University, Iwate, Japan*

Abstract

Video-based person re-identification (Re-ID) aims to match identical person sequences captured across non-overlapping surveillance areas. It is an essential yet challenging task to effectively embed spatial and temporal information into the video feature representation. For one thing, we observe that different frames in the video can provide complementary information for each other. Also, local features which is lost due to target occlusion or visual ambiguity in one frame can be supplemented by the same pedestrian part in other frames. For another thing, graph neural network enables the contextual interactions between relevant regional features. Therefore, we propose a novel sparse graph wavelet convolution neural network (SGWCNN) for video-based person Re-ID. Distinct from previous graph-based Re-ID methods, we exploit the weighted sparse graph to model the semantic relation among the local patches of pedestrians in the video. Each local patch in one frame can extract supplementary information from highly related patches in other frames. Moreover, to effectively solve the problems of short time occlusion and pedestrian misalignment, the graph wavelet convolution neural network is adopted for feature propaga-

*Corresponding author.
Email addresses: m025119302@sues.edu.cn (Yingmao Yao), xiaoyan.jiang@sues.edu.cn (Xiaoyan Jiang), hfujita@i-somet.org; hfujita-799@acm.org (Hamido Fujita), zjfang@sues.edu.cn (Zhijun Fang)

tion to refine regional features iteratively. Experiments and evaluation on three challenging benchmarks, that is, MARS, DukeMTMC-VideoReID, and iLIDS-VID, show that the proposed SGWCNN effectively improves the performance of video-based person re-identification.

Keywords: Video-based person re-identification, weighted sparse graph, graph wavelet convolution neural network

1. Introduction

Person Re-ID is usually regarded as an image retrieval problem, aiming to match the probe image with the gallery photos captured by different cameras. It has broad applications in the field of intelligent video surveillance and video analysis. Traditional image-based Re-ID methods, from hand-craft invariant features extraction [1] to end-to-end deep learning methods [2], encounter difficulties due to the limitation of information provided by single images. All these image-based models assume that one frame or several are selected in advance. However, when there are problems, such as, misaligned appearance, occlusion, and clutter background in single images, the performance of image-based models decline rapidly.

In contrast, video-based Re-ID models directly use video sequences as input and learn features in an end-to-end manner, which captures more information from multiple frames. Most existing video-based person Re-ID methods first extract image-level feature vectors from sequential frames [3, 4] and video-level feature vectors are generated by aggregating average or maximum pooling of image-level features, then compared these video-level features in a particular metric space. To handle problems caused by, such as, occlusion, illumination, viewing point, and pose variations in videos, recent works apply recurrent neural networks (RNN) [5], and 3D convolution (3D Conv) [6, 7] to learn the spatial-temporal dependency in an end-to-end way. Other methods adopt the attention mechanism [8, 9] to weight the importance of different frames or spatial loca-

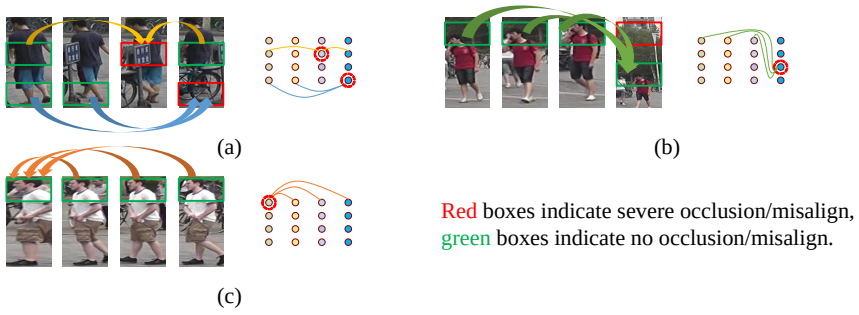


Figure 1: Three different video sequences in MARS dataset [10]. In our proposed method, each frame of a video sequence is horizontally (from the top to the bottom) segmented into S regions, known as patch features. Therefore, each patch feature represents the local feature of a human body. (a) due to camera viewing angle variance, the same body parts of the pedestrian in some video frames are temporarily occluded but reappear in other frames during movement. (b) pedestrian misalignment makes traditional CNNs integrate different body parts to represent one region. (c) the appearance of pedestrians in a video sequence usually do not change significantly.

tions to obtain a better representation. However, in challenging scenes, such as, occlusion and pose variations, these methods cannot effectively utilize the relationship between body parts across frames. Therefore, how to model the spatial-temporal relationship between regions across frame is a key issue to solve for video-based person re-identification.

First, as shown in Figure 1 (a), due to the change of camera viewing point, persons may be occluded for a short time during moving. In this case, using patches in other frames with high correlation of the current frame is helpful to recover the lost pedestrian patches. We explicitly exploit the temporal relations between local patch features across difference frames, which are helpful to provide complementary information for each other. Second, as shown in Figure 1 (b), person misalignment problem in the long-time period may cause CNNs to integrate non-corresponding body parts into the video-level feature representation. Hence, person patches should be spatially aligned for more robust feature

extraction. Third, as shown in Figure 1 (c), normally the person appearance may not change significantly in neighboring frames. Feature mapping for each pair of local patches in the video frames would cause a lot of information redundancy. Thus, we only need to carry out feature mapping for highly related local patches.

To solve the above issues and inspired by the success of graph convolution neural network (GCN) in image classification and target detection tasks [11], we propose to adopt GCN to model the spatial-temporal relations of different patches for person Re-ID. Different from the previous graph-based person Re-ID methods, such as, adaptive graph representation learning (AGRL) [12], and spatial-temporal graph convolutional network (STGCN) [13], we use weighted sparse graph instead of affinity graph to effectively capture the semantic relationship between highly correlated patches across frames, and reduce the mutual influence caused by unrelated body parts. In addition,

we use a spectral domain graph wavelet neural network [14] to implement efficient convolution on graph. Graph wavelet neural network distinguishes itself from spectral CNN by three desirable properties: (1) graph wavelet transform can be obtained by a fast algorithm without requiring Laplacian matrix eigendecomposition, thus it is efficient; (2) Graph wavelets are sparse, while eigenvectors of Laplacian matrix are dense. So, graph wavelet transform is much more efficient than graph Fourier transform; (3) Graph wavelets are localized in vertex domain, reflecting the information diffusion centered at each node [15]. This property eases the understanding of graph convolution defined by graph wavelets. Therefore, we propose a sparse graph wavelet convolution neural network (SGWCNN) to model the spatial-temporal relationship across different local patches of pedestrians in video sequences.

The main contributions are summarized as follows:

- We propose an adaptive weighted sparse graph generation approach, which

can efficiently capture the semantic relations between highly related body parts across frames, to alleviate occlusion, appearance misalignment, and information redundancy in videos.

- We present a unified end-to-end framework to model the spatial-temporal relation for different parts of the human body, mining more discriminative and robust information. Compared with traditional GCNs, SGWCNN contains less parameters and is more efficient.

- Experiments show that our proposed approach achieves state-of-the-art performance on three large-scale video-based person Re-ID datasets.

2. Related Work

2.1. Image-based Person Re-identification

Image-based person Re-ID has been extensively studied over the years. Early methods mainly focus on designing discriminative hand-crafted features [1]. With the success of CNNs, deep features learned from the network have replaced hand-crafted features. These networks can be divided into representation learning [2] and metric learning [16]. Representation learning uses the cross-entropy loss to learn discriminative identity features. Metric learning uses the triple loss to push features of the same persons and pull away features of different pedestrians. However, image-based methods lack modeling the temporal relationship of video sequences, which are inadequacy for video-based person Re-ID.

2.2. Video-based Person Re-identification

Most existing video-based person Re-ID methods use optical flow, RNNs, attention mechanism, or 3D CNNs to extract temporal information from videos. Yet, optical flow [4] is not robust enough for occlusion and noisy scenarios, and is time-consuming. RNN [5] has a limited effect on modeling temporal information in Re-ID task.

Li et al. [6] and Liao et al. [7] use 3D CNN to learn appearance and motion features together. But when the spatial misalignment exists, 3D convolution may mix the features of different body parts, which destroys the appearance representations of person. Spatiotemporal attention models [8] are proposed to solve the problem of occlusion in individual frames. Liu et al. [9] propose a non-local video attention network (NVAN), which improves the video representation by exploiting spatial and temporal characteristics in both low-level and high-level features. However, these methods do not fully consider the relationship of body parts across different frames. Therefore, although CNN-based models can extract local meaningful features that are shared with the entire datasets, they have problems in cases, such as, occlusion and misalignment of interframe images, which cause dramatic changes in object appearance.

2.3. Graph Model-based Methods

In recent years, graph neural networks (GNNs) have been successfully applied to many tasks in computer vision, such as, scene graph generation [17], point clouds classification, and action recognition [18]. There are some works apply GNNs to person Re-ID.

Nguyen et al. [19] ultilize graph convolutional networks to effectively construct and model the correlation between attributes and body parts with global features. Zhang et al. [20] propose a part-guided graph convolution network (PGCN) to learn the inter-local relationship and the intra-local relationship for feature representations. However, these video-based person Re-ID methods only use image-level feature, while the temporal information are not considered. Bao et al. [21] propose a masked graph attention network (MGAT), which combines the graph attention network (GAT) with a feature extractor to discover the relationship between frames and the variation of a region in the temporal domain. Similar to [21], an adaptive graph representation learning network (AGRL) [12] uses human key points to dynamically learn graph and model the inherent con-

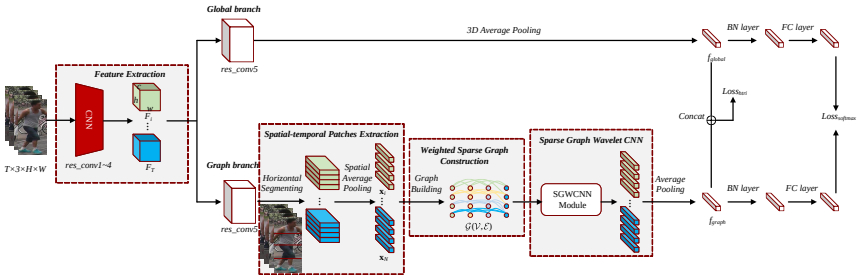


Figure 2: The proposed network architecture. 1) The ResNet50 backbone is adopted to extract the feature map for each frame of the input video. 2) In the global branch, the output feature of each frame is integrated into a video-level representation. 3) In the graph branch, the output patch features are integrated into a video-level representation. The green box and the blue box represent the feature map from different frames.

text relations in the image region. Yet, pose extraction requires additional prior information and extra computation, and it cannot perform an end-to-end training. Inspired by GCN [22], Yang et al. [13] adopt a unified framework named spatial-temporal graph convolution network (STGCN) that jointly considers the temporal and structural relations of all local patches in both video-level and image-level. Hence, the discriminability and spatial-temporal relations of patches can be learned to promote the video-based person Re-ID. But when all the patches in a video sequence are connected, the generated dense graph has a lot of information redundancy, and the patches from different body parts may not mutually promote. Instead of focusing on all pairs of patches, it is better to focus on a specific neighborhood.

3. The proposed method

3.1. Overview

The overall structure of the proposed network is shown in Figure 2. To balance accuracy and efficiency, the restricted random sampling (RRS) is used in the training period [12]. Given a video, we denote the sampling as $V =$

$\{I_t\}_{t=[1,T]}$, where T is the number of frames sampled from the video. For each frame of the video, the ResNet50 backbone is adopted to extract the feature map expressed as F:

$$F = \{F_1, F_2, ..., F_T\}, \tag{1}$$

where $F_i \in \mathbb{R}^{h \times w \times c}$ is the feature map of i-th frame in the video. And h, w, and c denote the height, width, and the channel number, respectively. Thus, each video sample consists of T frame image features.

In the global branch, 3D global average pooling is used for the feature maps extracted by the ResNet50 backbone. Besides, it aggregates the image features into the video representation $f_{global} \in \mathbb{R}^c$, where c is the dimension of feature channel.

In the graph branch, each feature map F_i is horizontally divided into a total S number of patches from the top to the bottom. Each patch is processed by the spatial average pooling to obtain its patch feature vector. Thus, the total number of patches is $N = T \cdot S$ for a video sample with T frames. We denote each patch of the video as $s_i \in S, i = 1, ..., N$, and the patch feature vector as $\mathbf{x}_i \in \mathbb{R}^c$, $i = 1, ..., N$. To capture the temporal-spatial relationship between each patch feature and its k highly related patch features, we construct a weighted sparse graph. Afterwards, in the SGWCNN module, a multi-layer graph wavelet convolution network is constructed to learn the context information of each patch feature and its neighbor nodes in the weighted sparse graph. Then an average pooling layer is used to aggregate the patch features into the video representation $f_{graph} \in \mathbb{R}^c$, where c is the dimension of feature channel.

The network is supervised by the combination of identification loss and triplet loss. Key modules are presented in the following.

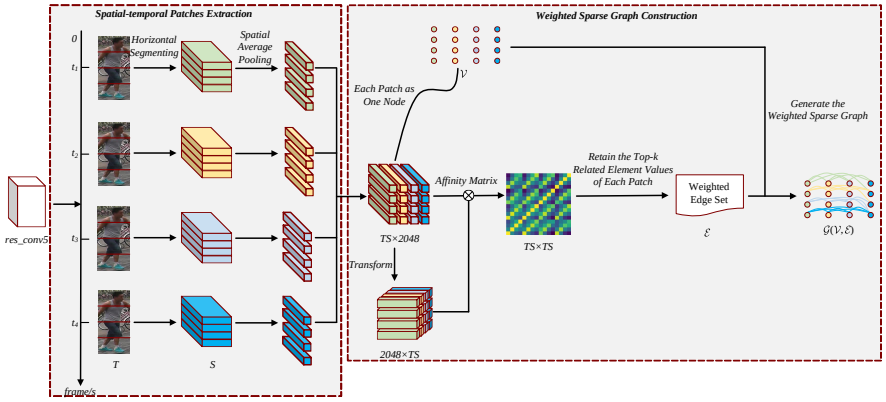


Figure 3: Illustration of the weighted sparse graph construction. Temporal-spatial patch features are generated by the spatial average pooling, where different colors represent different frames. In this module, each feature map is divided into S number of patches. Thus, we get $T \times S$ patches in total for a video sampling T frames. Each element in the affinity matrix is calculated to represent the pairwise relationship between patches, then record the top-k element values related to each patch in the weighted edge set.

3.2. Weighted Sparse Graph Construction

As mentioned, the body local patches of different frames in a video may provide complementary information. And, local patch features are useful in alleviating the problems caused by occlusion and noise. Thus, we first utilize the affinity adjacency matrix to capture the adjacency relation for each local patch pairs, after it a weighted sparse graph is constructed to model the context relationships of patch features.

Let $\mathcal{G}(\mathcal{V}, \mathcal{E})$ denotes the weighted sparse graph of N nodes, where $\mathcal{V} = \{\boldsymbol{v}_i\}_{i=1}^{N}$ is the vertex set of nodes, and the edges in $\mathcal{E}$ are used to represent the pairwise relations of patches. Each node $\boldsymbol{v}_i$ corresponds to a patch feature $\mathbf{x}_i$ in the frame, where $\boldsymbol{v}_i \in \mathbb{R}^c$. As shown in Figure 3, to define the edge $(\boldsymbol{v}_i, \boldsymbol{v}_j) \in \mathcal{E}$ in the graph, an affinity adjacency matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$ is firstly introduced, where each element represents a pairwise relation of patches. Then, we retain the top-k element values related to each patch to construct a weighted

sparse matrix $\tilde{\mathbf{A}} \in \mathbb{R}^{N \times N}$, and the k pairwise relations of patches are recorded in the weighted edge set. Finally, the graph is generated by the vertex set and the edge set.

3.2.1. Affinity Adjacency Matrix

To depict the spatial-temporal relationship between different patch features, an affinity adjacency matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$ is first constructed. Each element in the matrix represents the relation of two patch features. The features represented by two graph nodes $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$ are $\mathbf{x}_i$ and $\mathbf{x}_j$, respectively. Motivated by [23], the pairing relationship between every two patches is represented in the graph as follows:

$$\mathbf{r}_{i,j} = f_s(\mathbf{x}_i, \mathbf{x}_j) = {\mathbf{x}_i}^\top \mathbf{x}_j. \tag{2}$$

Hereto, the affinity adjacency of different temporal-spatial patches is obtained.

Then, normalization operation is performed to the adjacent matrix so that the sum of the element values in each row of the adjacency matrix is one:

$$\mathbf{A}_{i,j} = \frac{\mathbf{r}_{i,j}}{\sum_{j=1}^{N} |\mathbf{r}_{i,j}|}. \tag{3}$$

3.2.2. Weighted Sparse Graph

For each element in $\mathbf{A}$, we only retain the top-k related elements, resulting in a weighted sparse adjacency matrix $\tilde{\mathbf{A}}$. We have

$$\tilde{\mathbf{A}}_{i,j} = \begin{cases} \tilde{\mathbf{r}}_{i,j} & i \neq j \text{ and } \tilde{\mathbf{r}}_{i,j} \in \text{LARGE}\left(\tilde{\mathbf{A}}_{i,:}, k\right), \\ 0 & \text{otherwise}, \end{cases} \tag{4}$$

where $\tilde{\mathbf{r}}_{i,j}$ represents the normalized affinity score from node i to node j, and $\text{LARGE}\left(\tilde{\mathbf{A}}_{i,:}, k\right)$ means for the i-th row of matrix $\tilde{\mathbf{A}}$, only the top-k element values related to node i are retained. In this way, the number of edges on the graph decreases from N^2 to kN.

3.3. Sparse Graph Wavelet Convolution Neural Network

3.3.1. Graph Wavelet Convolution Definition

In traditional GCN, the graph Laplacian matrix $\mathcal{L}$ is defined as $\mathcal{L} = \mathbf{D} - \tilde{\mathbf{A}}$, where $\mathbf{D}$ is a diagonal degree matrix with $\mathbf{D}_{i,i} = \sum_j \tilde{\mathbf{A}}_{i,j}$. The normalized Laplacian matrix is $\mathbf{L} = \mathbf{I}_n - \mathbf{D}^{-1/2}\tilde{\mathbf{A}}\mathbf{D}^{-1/2}$, where $\mathbf{I}_n \in \mathbb{R}^{N\times N}$ is the identity matrix. Since $\mathcal{L}$ is a real symmetric matrix, it has a set of orthonormal eigenvectors $\boldsymbol{U} = (\boldsymbol{u}_1, \boldsymbol{u}_2, ..., \boldsymbol{u}_N)$, known as Laplacian eigenvector. The corresponding non-negative eigenvalues $\{\lambda_l\}_{l=1}^N$ of these eigenvectors are identified as the frequencies of graph.

Similar to graph Fourier transform, graph wavelet transform projects graph signal from the vertex domain into the spectral domain. Graph wavelet transform employs a set of wavelets $\psi_s = (\psi_{s1}, \psi_{s2}, ..., \psi_{sN})$ as bases, where each wavelet ψ_{si} corresponds to a signal on graph diffused away from node i, and s is a scaling parameter, constraining the range of the neighborhood. Mathematically,

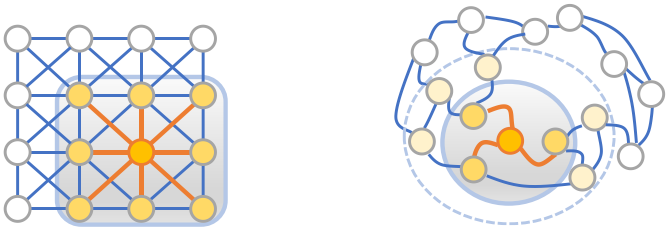
$$\psi_s = \boldsymbol{U}\boldsymbol{G}_s\boldsymbol{U}^\top, \tag{5}$$

where $\boldsymbol{U}$ is Laplacian eigenvectors, and the scaling matrix $\boldsymbol{G}_s = diag(g(s\lambda_1), ..., g(s\lambda_N))$ with $g(s\lambda_i) = e^{\lambda_i s}$.

Using graph wavelets as bases, graph transform of a signal $\mathbf{x}$ in graph is defined as $\hat{\mathbf{x}} = {\psi_s}^{-1}\mathbf{x}$, and the inverse graph wavelet transform is $\mathbf{x} = \psi_s\hat{\mathbf{x}}$. Note that ${\psi_s}^{-1}$ can be obtained by replacing the $g(s\lambda_i)$ in ψ_s with $g(-s\lambda_i)$ [15]. Graph wavelet transform, according to convolution theorem, offers us a way to define the graph wavelet convolution operator, denoted as $*_\mathcal{G}$ [14]. Denoting with $\boldsymbol{y}$ the convolution kernel, $*_\mathcal{G}$ is defined as:

$$\mathbf{x} *_\mathcal{G} \boldsymbol{y} = \psi_s((\psi_s^{-1}\boldsymbol{y}) \odot (\psi_s^{-1}\mathbf{x})), \tag{6}$$

where $\odot$ is the element-wise Hadamard product. We show the differences be-



(a) Traditional Convolution. Each pixel in an image is taken as a node where neighbors are determined by the filter size. The 2D convolution takes the weighted average of pixel values of the orange node along with its neighbors. The neighbors of a node are ordered and have a fixed size.

(b) Graph Wavelet Convolution. To get a hidden feature of the orange node, one simple solution of the graph wavelet convolution operation is to take the average value of the node features of the orange node along with its neighbors. Each edge have different weight and the neighbors of a node are variable in scale size.

Figure 4: Traditional Convolution vs. Graph Wavelet Convolution.

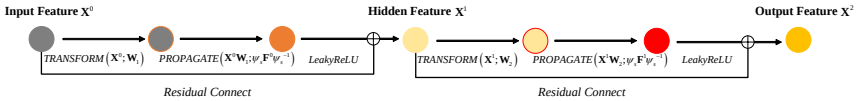


Figure 5: A two-layer SGWCNN module. Given a weighted sparse graph $\tilde{\mathbf{A}}$, the original spatial domain patch features first do a feature transform, and then the graph wavelet convolution operation is performed to complete the feature propagation. The operations are defined in Equations (7) and (8).

tween the traditional convolution and the graph wavelet convolution in Figure 4.

3.3.2. M-layer SGWCNN for Video-based Person Re-ID

In this paper, we construct a M-layer SGWCNN architecture, each layer in SGWCNN consists of feature transformation and graph convolution operation, as shown in Figure 5. For the n_{th}-layer ($1 \leq n \leq M$), we define feature transformation as:

$$\mathbf{X}^{n\prime} = \mathbf{X}^{n}\mathbf{W}, \tag{7}$$

where $\mathbf{X}^{n} \in \mathbb{R}^{N\times c}$ are the hidden features for all patches in the layer n, and c is the feature dimension; $\mathbf{X}^{0} \in \mathbb{R}^{N\times c}$ are the initial patch features obtained by the CNN backbone; $\mathbf{W} \in \mathbb{R}^{c\times c}$ is the parameter matrix to be learned. For the hidden features at layer $n+1$, we define graph convolution as:

$$\mathbf{X}^{n+1} = h(\psi_s \mathbf{F}^{n} {\psi_s}^{-1} \mathbf{X}^{n\prime}), \tag{8}$$

where ψ_s is the wavelet bases, and ${\psi_s}^{-1}$ is the graph wavelet transform matrix at scale s projecting signal in vertex domain into spectral domain. The input features are $\mathbf{X}^{n\prime} \in \mathbb{R}^{N\times c}$. $\mathbf{F}^{n}$ is the diagonal filter matrix learned in the spectral domain. And h is a non-linear activation function. In this paper, a LeakyReLU (with negative input slope $\alpha = 0.1$) is used as non-linear activation function. In addition, we use residual connection [24] to make our network easier to optimize,

$$\mathbf{X}^{n+1} := \mathbf{X}^{n+1} + \mathbf{X}^{n}, 1 \leq n \leq M-1, \tag{9}$$

after M layers graph convolution, the output of each video is $\mathbf{X}^{M} \in \mathbb{R}^{N\times c}$. Finally, we use the average pooling operation on $\mathbf{X}^{M}$. Therefore, for each video, we obtain its graph branch feature $f_{graph} \in \mathbb{R}^{c}$.

3.4. Loss Functions

In this paper, we utilize both the cross-entropy loss and the batch hard triplet loss to train the model through the combination the discriminative learning and the metric learning.

The cross-entropy loss is used to calculate the classification error among the identities, which is formulated as follows:

$$L_{softmax} = -\frac{1}{P\cdot K}\sum_{i=1}^{P}\sum_{a=1}^{K} log\frac{exp\left({\mathbf{W}_{y_{i,a}}}^{\top}\mathbf{x}_{i,a}\right)}{\sum_{m=1}^{P\cdot K} exp\left({\mathbf{W}_{m}}^{\top}\mathbf{x}_{i,a}\right)}, \tag{10}$$

where P and K are the number of identities and sampled images for each identity, respectively. $\mathbf{x}_{i,a}$, $\mathbf{x}_{i,p}$, and $\mathbf{x}_{i,n}$ are the features extracted from anchor, positive and negative samples, respectively. And $\mathbf{y}_{i,a}$ is the ground truth identity of the sample $\{i, a\}$.

The triplet loss is first proposed in [16], which is used to teach the network to push together features of the same person and pull away features of different people. We use the triplet loss with hard mining [12] as follows:

$$L_{triplet} = \sum_{i=1}^{P}\sum_{a=1}^{K} ln(1 + exp(\overbrace{\max_{p=1,\dots,K} D(\mathbf{x}_{i,a}, \mathbf{x}_{i,p})}^{hardest\ positive} - \underbrace{\min_{\substack{n=1,\dots,K \\ j=1,\dots,P \\ j\neq i}} D(\mathbf{x}_{i,a}, \mathbf{x}_{j,n})}_{hardest\ negative})), \tag{11}$$

where $D(\cdot)$ is the L2-norm distance of two feature vectors.

For the final softmax loss $L'_{softmax}$, we individually compute each softmax loss of the two extracted features:

$$L'_{softmax} = L^{global}_{softmax} + L^{graph}_{softmax}. \tag{12}$$

For the batch-hard triplet loss $L_{triplet}$, as shown in Figure 2, we concatenate the two extracted features f_{global} and f_{graph} as the final feature, which can be written as $f_{all} = [f_{global}, f_{graph}]$, where $[\cdot]$ means concatenation. We use the feature f_{all} to compute the triplet loss.

Finally, the total loss L_{total} is the summation of the above two loss items:

$$L_{total} = L'_{softmax} + \lambda L_{triplet}, \tag{13}$$

where λ refers to the trade-off parameter between the losses (we use $\lambda = 1$, as per [12]).

Figure 6: Example image sequences of the same person appearing in different camera views from three datasets. There are six camera views on (a), four camera views on (b), and two camera views on (c).

4. Experiments

The proposed model is evaluated on three large scale video-based person Re-ID datasets, that is, MARS [10], DukeMTMC-VideoReID [25], and iLIDS-VID [26]. We show example image sequences in Figure 6.

4.1. Datasets and Evaluation Protocols

MARS. It is the largest video-based person Re-ID benchmark with 17,503 sequences of 1261 identities and 3,248 distractor sequences generated by DPM detector and GMMCP tracker. The dataset is captured by 6 cameras and has 13.2 sequences on average. There are 3,248 distracter sequences in MARS, which

makes it more challenging.

DukeMTMC-VideoReID. It is another large-scale benchmark dataset for video-based person Re-ID, which is derived from the DukeMTMC dataset and re-organized by Wu et al. [27]. DukeMTMC-VideoReID contains 4,832 tracklets of 1,812 identities. It is divided into 702, 702, and 408 identities for training, testing, and distraction, respectively. There are a total number of 2,196 sequences containing 369,656 frames used for training, and 2,636 sequences containing 445,764 frames used for testing and dispersion. Each tracklet has 168 frames on average, and the bounding boxes are annotated manually.

iLIDS-VID. It contains 600 image sequences of 300 people. Everyone has two videos, ranging in length from 23 to 192 frames, with an average length of 73 frames.

Evaluation Protocols. Cumulative Matching Characteristic (CMC) accuracy and the mean average precision (mAP) are used to evaluate the performance of the proposed method. CMC judges the ranking capabilities of the Re-ID model and mAP reflects the true ranking results when multiple corresponding ground-truth sequences exist. In this paper, rank-1, rank-5, and rank-20 are used as the CMC evaluation standard. For MARS and DukeMTMC-VideoReID datasets, both CMC and mAP are reported. For iLIDS-VID dataset, we follow the evaluation protocol used in [26]. The dataset is divided into two parts for training and testing. The final accuracy is the average of "10-fold cross validation". Only CMC accuracy is reported in iLIDS-VID since CMC and mAP are equal for this dataset.

4.2. Implementation Details

The experiments are implemented with Pytorch and with one NVIDIA RTX 3090 GPU. ResNet50 [24] is used as the backbone, which is firstly pre-trained on ImageNet. The last stride of ResNet50 is set to be 1. The input images are all resized to 256×128 with random horizontal flips [9] for data augmentation.

In the training stage, a restricted random sampling strategy is employed to randomly sample $T = 8$ frames from every video, which are grouped into tracklet. The parameters are updated by employing Adam [28], with an initial learning rate of 3×10^{-4}, and the weight decay of 5×10^{-4}. According to [9], eight identities are sampled, and each identity has four tracklets to form a batch size of $8 \times 4 \times 8 = 256$ images. We train the network for 240 epochs in total, and the learning rate decays by 10 every 60 epochs. For batch triplet loss, we set $P = 8$ and $K = 4$ for all our experiments. So, the batch size is set to 32. For the parameters of the graph branch, in the weighted sparse graph construction, we set the number of adjacent edges k of each node to 3. In the SGWCNN module, we set the number of layers M to 2, the scale parameter s to 1, and the number of patches S to 4. We concatenate the features of global branch and graph branch, then use the cosine similarity as the distance metric to match the final video representations.

We follow the standard configuration in the testing stage for accurate evaluation. A video containing T_v frames is split into T chunks firstly. The dense sampling strategy samples all frames in the video into a list of sequences [12], each sequence contains T frames and the batch size is set to 1.

4.3. Comparison with State-of-the-Art Methods

We compare our approach with state-of-the-art methods on MARS, DukeMTMC-VideoReID, and iLIDS-VID datasets. As we can see from Table 1, the accuracy of rank-1 and mAP of SGWCNN on MARS achieves the best performance with 90.0% and 85.7%, respectively. In DukeMTMC-VideoReID dataset, only mAP achieves the best performance with 95.9%, while rank-1 score is 96.3%. We think this is related to the fact that each identity has fewer video sequences, and the image sizes of different datasets have an influence on the patch features resolution, when the accuracy is high, the bottleneck is encountered. And the accuracy of rank-1 of SGWCNN on iLIDS-VID achieves the best performance

Table 1: Comparison with state-of-the-art video-based person re-id methods on MARS, DukeMTMC-VideoReID, and iLIDS-VID datasets. Results are shown regarding to matching rates (%) at rank = 1, 5, 20 on all datasets and mAP on MARS and DukeMTMC-VideoReID datasets. The MGAT, AGRL, and STGCN are graph-based methods.

Method	MARS				DukeMTMC-VideoReID				iLIDS-VID		
	mAP	rank1	rank5	rank20	mAP	rank1	rank5	rank20	rank1	rank5	rank20
CNN+XQDA [10]	47.6	65.3	82.0	89.0	-	-	-	-	53.0	81.4	95.1
DAM [4]	57.7	74.7	87.0	93.1	-	-	-	-	77.3	94.0	**100**
HCRN [5]	67.1	73.9	91.6	97.3	-	-	-	-	71.3	91.4	99.2
M3D [6]	78.6	86.3	-	-	92.3	94.6	-	-	81.3	-	-
STA [8]	80.8	86.3	95.7	98.1	94.9	96.2	99.3	99.6	-	-	-
STE-NVAN [9]	81.2	88.9	-	-	93.5	95.2	-	-	-	-	-
VRSTC [29]	82.3	88.5	96.5	-	93.5	95.0	99.1	-	83.4	95.5	99.5
B-BOT [30]	82.9	88.6	96.2	98.0	-	-	-	-	-	-	-
DPRM [31]	83.0	89.0	96.6	98.3	95.6	97.1	99.4	**100**	-	-	-
ResVKD-50bam [32]	83.1	89.4	96.8	-	93.5	95.2	98.6	-	-	-	-
TSF [33]	85.2	87.1	96.8	-	-	-	-	-	87.7	-	-
MGAT [21]	71.8	81.1	92.2	97.7	-	-	-	-	80.3	94.7	99.5
AGRL [12]	81.1	89.8	96.6	97.8	94.2	96.7	99.2	99.7	83.7	95.4	99.5
STGCN [13]	83.7	89.9	96.4	98.2	95.7	**97.3**	99.3	99.7	-	-	-
SGWCNN(Ours)	**85.7**	**90.0**	**97.0**	**98.4**	**95.9**	96.3	**99.4**	99.9	**87.8**	**96.0**	99.3

with 87.8%.

Compared with the existing attention-based methods (including STA [8], STE-NVAN [9], B-BOT [30], DPRM [31], TSF [33]) do not make full use of the complementary relations from the same body parts, thus they lack discriminative information. The 3D convolution method M3D [6] is sensitive to spatial misalignment and is computationally expensive. The generating-based method VRSTC [29] try to recover the occluded area, but this method is restricted by the effect of the generator. The method based on knowledge distillation (ResVKD-50bam [32]) requires training of two networks, the teacher educates a student who observes fewer views, but the effect of student network largely depends on the performance of the teacher network. Compared with other graph-based methods (including MGAT [21], AGRL [12], STGCN [13]), in [21], the attention mechanism is used to reveal the relative importance between two image-level features, thus they ignored the interaction between two local patches

on the time series. The pose estimation in [12] is separate from the entire network framework, which may cause sub-optimal result and the pose alignment is sensitive to the quality of pose estimation. In [13], the pairwise relations of patches are directly modeled by a dense affinity graph, the graph has both redundant information and may make two completely unrelated local patches influence each other.

Compared with other methods, our proposed approach enables each patch to make full use of the complementary information of the k patch features that are highly related. Graph models can better solve the problems of occlusion and misalignment, and the network can be trained in an end-to-end way.

4.4. Ablation Study

To verify the impact of the sparse graph wavelet convolution module, we first replaced the SGWCNN layer with a feature transformation layer, then perform training and testing under the same experimental setting. We conduct ablation study on MARS dataset. Specially, considering a one-layer SGWCNN, which can be described by Equation (7) and (8). On the other hand, the formulation of the feature transformation network (FTN) can be written as $Y = h(\mathbf{XW})$, where $\mathbf{X}$ is the input and $\mathbf{W}$ is the parameters matrix of the FTN layer. Compared with SGWCNN, the FTN can be viewed as removing the sparse graph wavelet matrix $\psi_s \mathbf{F} {\psi_s}^{-1}$.

As shown in Table 2, the baseline contains only the ResNet50 backbone and 3D average pooling, and is supervised by the combination of $L_{softmax}^{global}$ and $L_{triplet}^{global}$. The mAP and rank-1 accuracy of the baseline is 84.2% and 88.7%, respectively. We find that "global+FTN" is limited in improving the rank-1 accuracy. It is because that using the feature transformation layer (Equation (7)) alone cannot model the relations of different patch features, such method cannot further mine the spatial-temporal information in a video. Compared with "global+FTN", "global+SGWCNN" improves the mAP and rank-1 accu-

Table 2: Ablation study on MARS dataset. We present matching rates (%) at rank = 1, 5, 20, and mAP. global+FTN: the addition of a FTN. global+SGWCNN: the addition of a 1-layer Sparse Graph Wavelet CNN. global+SGWCNN (2 layers): the addition of a 2-layer Sparse Graph Wavelet CNN. The baseline consists of feature extractor and 3D average pooling.

Method	MARS			
	mAP	rank1	rank5	rank20
Baseline (only global branch)	84.2	88.7	96.0	97.7
global+FTN	85.2	88.5	96.4	97.8
global+SGWCNN	**85.8**	89.7	96.3	98.1
global+SGWCNN (2 layers)	85.7	**90.0**	**97.0**	**98.4**

Table 3: Comparison with graph-based person re-id methods on MARS, DukeMTMC-VideoReID, and iLIDS-VID datasets. We present matching rates (%) at rank = 1 on all datasets and mAP on MARS and DukeMTMC-VideoReID datasets. In addition, we present the number of parameters and the running speed of the models.

Method	MARS			DukeMTMC-VideoReID			iLIDS-VID		Speed
	mAP	rank1	Params	mAP	rank1	Params	rank1	Params	
baseline+AGRL [12]	85.3	89.6	49.4MB	95.6	95.9	49.8MB	87.1	47.5MB	**304FPS**
baseline+STGCN [13]	85.2	89.7	61.7MB	95.4	95.7	62.2MB	83.2	58.8MB	96FPS
baseline+SGWCNN	**85.7**	**90.0**	**41.0MB**	**95.9**	**96.3**	**41.3MB**	**87.8**	**39.1MB**	128FPS

racy by 0.6% and 1.2%, respectively. The improvement in accuracy proves the effectiveness of sparse graph wavelet (Equation (8)). This is because the graph wavelet convolution layer utilizes complementary information between different patches to extract more discriminative and robust person features. Furthermore, our proposed "global+SGWCNN (2 layers)" gains 1.5% and 1.3% higher scores in mAP and rank-1 accuracy than the baseline, respectively. Finally, we can obtain 85.7% in mAP and 90.0% in rank-1 accuracy.

As shown in Table 3, we compare SGWCNN with the adaptive graph representation learning (AGRL) and the spatial-temporal graph convolution network (STGCN). The rank-1 accuracy of "Baseline+SGWCNN" on all datasets has reached the highest, and the MAP accuracy on MARS and DukeMTMC-VideoReID has also reached the highest. These results show that the weighted sparse graph extracts complementary identity features from highly correlated

Table 4: Statistical comparison of our sparse graph wavelet transform and the traditional dense graph Fourier transform on a video sequence.

	Statistical Property	Sparse Graph Wavelet Transform	Dense Graph Fourier Transform
Transform Matrix	Density of Non-zero Elements	18.36%	100%
	Number of Non-zero Elements	188	1,024
Projected Signal	Density of Non-zero Elements	36.52%	100%
	Number of Non-zero Elements	374	1,024

patches. Compared with the previous methods that require extracting pair-by-pair relation of patches [12, 13], the spatial and temporal information mined by our method is more discriminative. Meanwhile, the spectral domain graph wavelet convolution is sparse and localized in vertex domain, thus the results are better and more explainable. Furthermore, the number of parameters of "Baseline+SGWCNN" on all datasets is the minimum.

In addition, we find that the speed of "Baseline+SGWCNN" is faster than that of "Baseline+STGCN" on all datasets, because the STGCN establishes a GCN for each frame in image sequences, which is less efficient. However, the speed of "Baseline+SGWCNN" is slower than "Baseline+AGRL", this could be caused by the operation speed of graph wavelet in the spectral domain, which is slower than that of graph propagation method in the spatial domain.

We also analyze the number of non-zero elements in the weighted graph of a video sequence between the wavelet transform of the sparse graph and the Fourier transform of the dense graph in Equation (5). As shown in Table 4, our sparse graph wavelet transform has fewer parameters, which enables the network to focus on more closely connected neighborhoods while achieving better performance.

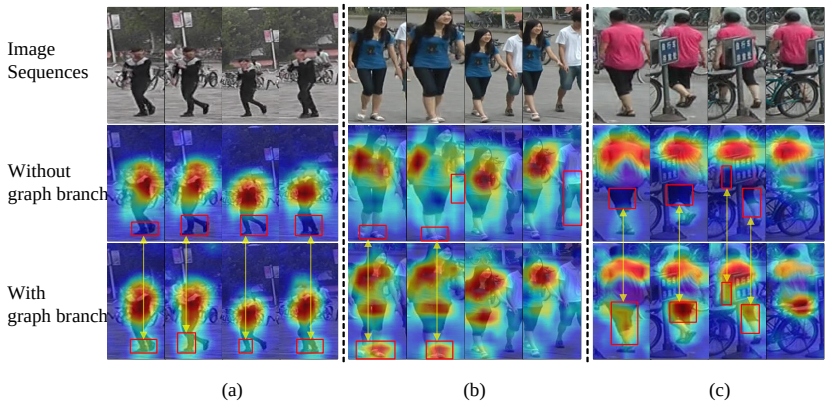


Figure 7: Visualization of the class activation maps (CAMs). The first row shows the original image sequences from MARS. In the other rows, the class activation maps for the baseline model and our proposed method are provided. Compared with the baseline model without graph learning, we find out that our proposed method is robust to the misaligned appearance, clutter background, and occlusion.

4.5. Visualization

4.5.1. Visualization of Class Activation Maps

We visualize the class activation maps in Grad-CAM [34], which is usually used in computer vision problems for visual explanation of the network. We can observe that the class activation maps of different frames show highlights in the same discriminative area. The misaligned appearance cases are shown in Figure 7 (a). We can see that model with graph branch focuses the attention area on the whole human body, but the baseline model only focuses on the local parts of human body. Figure 7 (b) shows cases with background clutter. When the distractions appear in the background, the attention area of the identity deviates to the background area, while for the model with graph branch, the influence of distractions is weakened. Figure 7 (c) shows the occlusion cases where the identity is partially occluded by the barriers. From the visualization can see that our proposed method utilizes the spatial-temporal relations of patches to

mine more discriminative cues, however, these cues are missing from the baseline model.

4.5.2. Retrieval Results

As illustrated in Figure 8, we provide the retrieval results of the baseline and our proposed method on MARS dataset. Videos capturing the same person under three different camera angles were used as the query samples. We can see that only one of the rank-5 results of the three videos is mismatched by our proposed method. In the first video, the rank-4 and rank-5 results of the baseline model are disturbed by occlusion and similar appearance. In our proposed method, the appearance misalignment and occlusion do not affect the retrieval results in rank-4 and rank-5. In the second video, the rank-2 and rank-3 results of the baseline model are disturbed by appearance misalignment. However, in our proposed method, this situation has improved. Therefore, the retrieval results prove that our proposed approach indeed alleviate the problems of occlusion, similar appearances, and misalignment.

4.6. Further Analysis

In our proposed method, S indicates the number of feature map patches, k indicates the number of adjacent edges retained by each sparse graph node. The scaling parameter of the graph wavelet transform is s, and the number of SGWCNN layers is M. As shown in Table 5, we conduct experiments to investigate the effect of different parameter settings.

The impact of the number of layers in SGWCNN. We fix $S = 4$, $k = 3$, and $s = 1$, then evaluate the results when $M = 1, 2, 3$. As shown in Figure 9 (a), we can see that the best Rank-1 is 90.0% when $M = 2$, and the best mAP is 85.8% when $M = 1$. If the layer is too shallow, the SGWCNN cannot effectively propagate the node information to the entire data graph. But when the SGWCNN is too deep, it also brings potential over-smoothing [35].

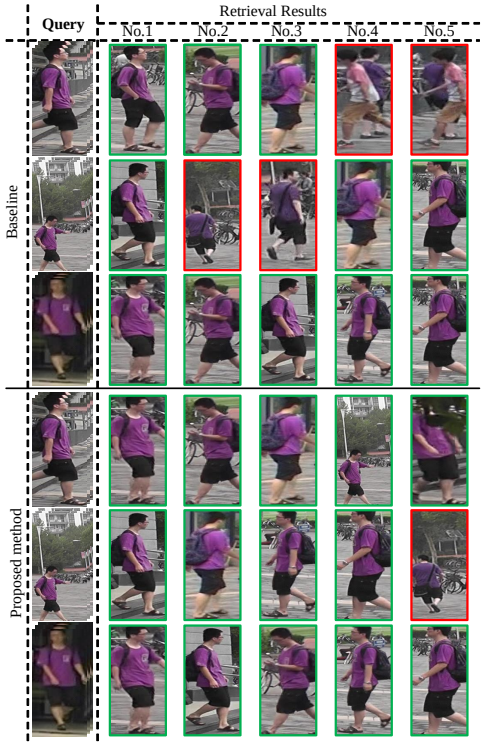


Figure 8: Comparison of rank-5 of our proposed method and the baseline model. In each row, the images represent the videos in the gallery, which are captured from different viewpoints. Images with green boxes indicate right matches with the query and with red boxes indicate wrong matchings.

The impact of the number of retained edges for each node. Similarly, we fix $S = 4$, $M = 2$, and $s = 1$, then evaluate the results when $k = 1, 3, 5$. As shown in Figure 9 (b), the model obtains the best performance when $k = 3$. It indicates that graph should not be too sparse or dense.

The impact of the value of scaling parameter. We fix $S = 4$, $M = 2$, and $k = 3$, then evaluate the results when $s = 1, 2$. As shown in Figure 9 (c), the model has the best performance when $s = 1$. The scaling parameter affects the neighbor distance of the node information diffusion. In our experiments, the best performance can be obtained by fusing the vertex feature and its one-hop

Table 5: The performance of SGWCNN on MARS using different parameter settings. 'Patches' represents the number of patches in each feature map. 'Top-k' represents the number of adjacent edges retained by nodes in the sparse graph. 'Scale' represents the scaling parameter of the graph wavelet transform. 'Layers' represents the number of layers of SGWCNN.

Patches	Top-k	Scale	Layers	mAP	rank1
2	3	1	2	85.4	89.1
4	3	1	2	85.7	**90.0**
8	3	1	2	85.2	89.1
4	1	1	2	85.5	89.7
4	5	1	2	85.0	89.0
4	3	2	2	85.6	89.4
4	3	1	1	**85.8**	89.7
4	3	1	3	85.2	88.4

neighbor nodes.

The impact of the number of patches. We fix $M = 2$, $k = 3$, and $s = 1$, then evaluate the results when $S = 2, 4, 8$. As shown in Figure 9 (d), the model has the best performance when $S = 4$. It indicates that number of local patches should be moderate.

5. Conclusion

In this paper, we propose a SGWCNN to model the spatial-temporal relationship between different local patches of inter-frame for video-based person re-identification. Meanwhile, we present an approach to adaptively generate weighted sparse graphs, in which each patch feature integrates complementary information from the highly related patches in the video sequence. In this way, we could enhance the representation capacity of video features to alleviate the problems existing in patches, such as, misaligned appearance, occlusion, and clutter background. Extensive evaluations have been carried out. Experimental results on three public datasets, that is, MARS, DukeMTMC-VideoReID, and iLIDS-VID, show the high re-identification performance of SGWCNN and the

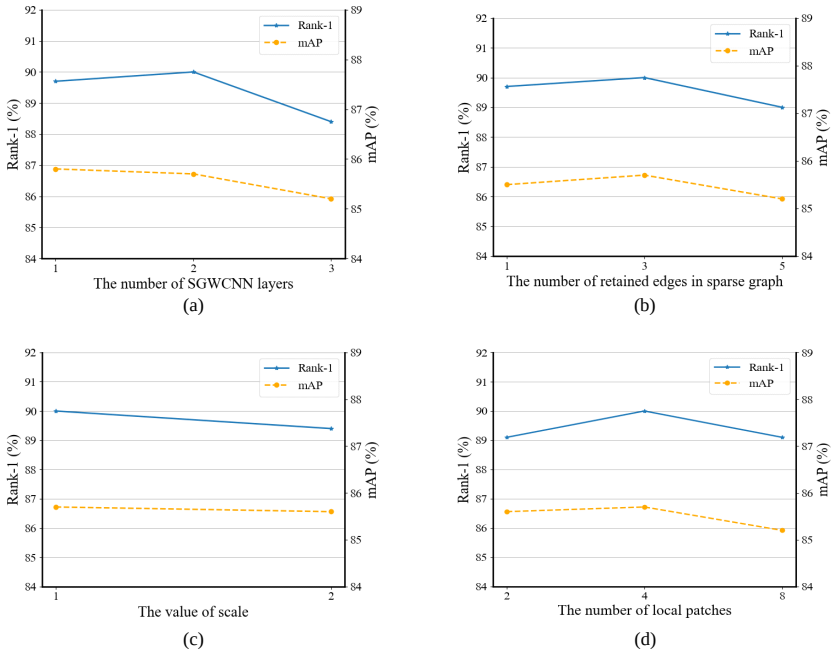


Figure 9: (a) Analysis on the number of SGWCNN layers M. (b) Analysis on the number of retained edges k for each node. (c) Analysis on the value of scale s. (d) Analysis on the number of local patches S. We carry out these experiments on the MARS datasets.

importance of learning spatial-temporal relationships between local patches.

Although the proposed model achieves good performance, the sampling method for person patch features generated by horizontal segmentation may contain invalid background regions. In the future work, we plan to make the sampled patch features more discriminable and study how to reduce the influence of noisy patches on graph models. For example, the key points of the human body can be used to define the patch features, so that the body structural information contained in each patch feature is more explicit. Moreover, the topology of graph model can be optimized by introducing the related works of graph pooling to alleviate the interference of noise nodes.

6. Acknowledgments

This work is supported by the following projects: National Natural Science Foundation of China (NSFC) Nr.: U2033218, 61831018.

References

[1] C. Zhao, X. Wang, D. Miao, H. Wang, W. Zheng, Y. Xu, D. Zhang, Maximal granularity structure and generalized multi-view discriminant analysis for person re-identification, Pattern Recognition (PR) Vol: 79 (2018) pp: 79–96.

[2] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Z. Hu, C. Yan, Y. Yang, Improving person re-identification by attribute and identity learning, Pattern Recognition (PR) Vol: 95 (2019) pp: 151–161.

[3] G. Chen, J. Lu, M. Yang, J. Zhou, Learning recurrent 3d attention for video-based person re-identification, IEEE Transactions on Image Processing (TIP) Vol: 29 (2020) pp: 6963–6976.

[4] J. Meng, A. Wu, W.-S. Zheng, Deep asymmetric video-based person re-identification, Pattern Recognition (PR) Vol: 93 (2019) pp: 430–441.

[5] L. Cheng, X.-Y. Jing, X. Zhu, F. Ma, C.-H. Hu, Z. Cai, F. Qi, Scale-fusion framework for improving video-based person re-identification performance, Neural Computing and Applications (NCA) Vol: 32 (16) (2020) pp: 12841–12858.

[6] J. Li, S. Zhang, T. Huang, Multi-scale temporal cues learning for video person re-identification, IEEE Transactions on Image Processing (TIP) Vol: 29 (2020) pp: 4461–4473.

[7] S. Li, H. Yu, H. Hu, Appearance and motion enhancement for video-based person re-identification, in: AAAI Conference on Artificial Intelligence (AAAI), 2020, pp. 11394–11401.

[8] Y. Fu, X. Wang, Y. Wei, T. Huang, Sta: Spatial-temporal attention for large-scale video-based person re-identification, in: AAAI Conference on Artificial Intelligence (AAAI), 2019, pp. 8287–8294.

[9] C.-T. Liu, C.-W. Wu, Y.-C. F. Wang, S.-Y. Chien, Spatially and temporally efficient non-local attention network for video-based person re-identification, in: British Machine Vision Conference (BMVC), 2019.

[10] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, Q. Tian, Mars: A video benchmark for large-scale person re-identification, in: European Conference on Computer Vision (ECCV), 2016, pp. 868–884.

[11] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, S. Y. Philip, A comprehensive survey on graph neural networks, IEEE Transactions on Neural Networks and Learning Systems (TNNLS) Vol: 32 (1) (2021) pp: 4–24.

[12] Y. Wu, O. E. F. Bourahla, X. Li, F. Wu, Q. Tian, X. Zhou, Adaptive graph representation learning for video person re-identification, IEEE Transactions on Image Processing (TIP) Vol: 29 (2020) pp: 8821–8830.

[13] J. Yang, W.-S. Zheng, Q. Yang, Y.-C. Chen, Q. Tian, Spatial-temporal graph convolutional network for video-based person re-identification, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3289–3299.

[14] B. Xu, H. Shen, Q. Cao, Y. Qiu, X. Cheng, Graph wavelet neural network, in: International Conference on Learning Representations (ICLR), 2018.

[15] J. Wang, Z. Deng, A deep graph wavelet convolutional neural network for

semi-supervised node classification, in: International Joint Conference on Neural Networks (IJCNN), 2021, pp. 1–8.

[16] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: A unified embedding for face recognition and clustering, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 815–823.

[17] A.-A. Liu, H. Tian, N. Xu, W. Nie, Y. Zhang, M. Kankanhalli, Toward region-aware attention learning for scene graph generation, IEEE Transactions on Neural Networks and Learning Systems (TNNLS) (2021) pp: 1–12.

[18] Z. Liu, H. Zhang, Z. Chen, Z. Wang, W. Ouyang, Disentangling and unifying graph convolutions for skeleton-based action recognition, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 143–152.

[19] B. X. Nguyen, B. D. Nguyen, T. Do, E. Tjiputra, Q. D. Tran, A. Nguyen, Graph-based person signature for person re-identifications, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3492–3501.

[20] Z. Zhang, H. Zhang, S. Liu, Y. Xie, T. S. Durrani, Part-guided graph convolution networks for person re-identification, Pattern Recognition (PR) Vol: 120 (2021) pp: 108155.

[21] L. Bao, B. Ma, H. Chang, X. Chen, Masked graph attention network for person re-identification, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[22] T. N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, in: International Conference on Learning Representations (ICLR), 2017.

[23] Z. Zhang, C. Lan, W. Zeng, X. Jin, Z. Chen, Relation-aware global attention for person re-identification, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3186–3195.

[24] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[25] E. Ristani, F. Solera, R. Zou, R. Cucchiara, C. Tomasi, Performance measures and a data set for multi-target, multi-camera tracking, in: European Conference on Computer Vision (ECCV), 2016, pp. 17–35.

[26] T. Wang, S. Gong, X. Zhu, S. Wang, Person re-identification by video ranking, in: European Conference on Computer Vision (ECCV), 2014, pp. 688–703.

[27] 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: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5177–5186.

[28] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: International Conference on Learning Representations (ICLR), 2015.

[29] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, X. Chen, Vrstc: Occlusion-free video person re-identification, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7183–7192.

[30] P. Pathak, A. E. Eshratifar, M. Gormish, Video person re-id: Fantastic techniques and where to find them (student abstract), in: AAAI Conference on Artificial Intelligence (AAAI), 2020, pp. 13893–13894.

[31] X. Yang, L. Liu, N. Wang, X. Gao, A two-stream dynamic pyramid repre-

sentation model for video-based person re-identification, IEEE Transactions on Image Processing (TIP) Vol: 30 (2021) pp: 6266–6276.

[32] A. Porrello, L. Bergamini, S. Calderara, Robust re-identification by multiple views knowledge distillation, in: European Conference on Computer Vision (ECCV), 2020, pp. 93–110.

[33] X. Jiang, Y. Gong, X. Guo, Q. Yang, F. Huang, W.-S. Zheng, F. Zheng, X. Sun, Rethinking temporal fusion for video-based person re-identification on semantic and time aspect, in: AAAI Conference on Artificial Intelligence (AAAI), 2020, pp. 11133–11140.

[34] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: IEEE International Conference on Computer Vision (ICCV), 2017, pp. 618–626.

[35] M. Chen, Z. Wei, Z. Huang, B. Ding, Y. Li, Simple and deep graph convolutional networks, in: International Conference on Machine Learning (ICML), 2020, pp. 1725–1735.