IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 19, 2020, accepted February 18, 2021, date of publication March 1, 2021, date of current version March 5, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3062967

Pose-Guided Multi-Scale Structural Relationship Learning for Video-Based Pedestrian Re-Identification

DAN WEI, XIAOQIANG HU, ZIYANG WANG, JIANGLIN SHEN, AND HONGJUAN REN
School of Mechanical and Automotive Engineering, Shanghai University of Engineering Science, Shanghai 201620, China

Corresponding author: Dan Wei (weiweidandan@163.com)

This work was supported by the National Science Foundation of China under Grant 51805312.

ABSTRACT How to extract discriminative features from redundant video information is a key issue for video pedestrian re-identification. Factors such as occlusion, perspective, and posture changes in complex environments pose severe challenges to pedestrian re-identification based on local methods. In this paper, a posture-guided multi-scale structural relationship learning pedestrian re-identification method is proposed. The purpose is to analyze the video sequence of pedestrians based on the reference pose and the pose alignment model, and extract the sample frame with the highest image quality and the most complete spatial information in the reference pose. The method based on posture guidance can more accurately eliminate the interference of background, occlusion and perspective factors. To further explore the potential relationship between local regions, this paper calculates the relationship matrix between the local regions based on the relationship model to further calculate the relationship weight, and the graph convolutional network based on the relationship weight learns the structural relationship feature of multi-scale regions. The input of the graph convolutional network is a local region divided by a multi-scale method, and the output is a pose-guided multi-scale structural relationship feature. The experimental results on three public datasets show that the proposed method performs favorably against state-of-the-art methods.

INDEX TERMS Pedestrian re-identification, relationship model, graph convolutional network, multi-scale structure relationship.

I. INTRODUCTION

Pedestrian re-identification is a technology to determine whether there is a specific pedestrian in an image or video sequence. Given an image or video sequence of a specific pedestrian, perform pedestrian image matching in the case of cross-monitoring equipment. At present, it is widely used in intelligent video surveillance, intelligent security and other fields. It has great application value for finding missing persons, tracing criminals, security, and monitoring. In recent years, it has attracted widespread attention in the field of computer vision. The methods to solve this problem are roughly divided into three categories, feature representation [1], metric learning [2], [3], and deep learning [4].

Extracting robust feature representations is the key to pedestrian re-identification. Compared with single-frame images, video sequences contain rich spatial information and timing information. Reasonable and efficient use of video information is an important criterion for algorithm performance. Blind use of all video sequences will cause immeasurable cost waste. After careful observation of multiple pedestrian video data sets, it is found that each pedestrian's video sequence contains a lot of redundant information. Moreover, the number of video frames of pedestrians is not the same, and while there is occlusion, it also includes changes in multiple viewing angles and postures. Aiming at the problem of information redundancy in video sequences, Wu *et al.* proposed an adaptive graph representation learning method, which uses posture alignment connections and features affinity connections to construct an adaptive structure-perception adjacency graph. In this method, the author uses a restricted random sampling method to select sample frames from the video, and for each frame of the image, a horizontal division method is used to construct

The associate editor coordinating the review of this manuscript and approving it for publication was Mingbo Zhao.

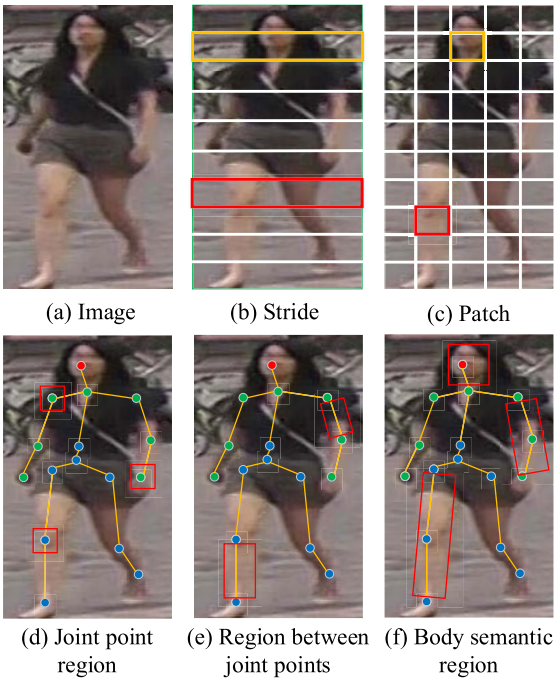


FIGURE 1. Multi-scale regions (d)-(f) based on posture guidance are used to accurately extract the features of the target region in this paper. Both horizontal strips (b) and patch (c) contain a lot of background noise.

local features, as shown in Fig. 1 (b). Randomly selected sampling frames will have occlusion, low image resolution, etc., and the horizontal division of the image will destroy the structural relationship between various parts of the human body, and will also introduce additional background noise [5]. Yang *et al.* proposed a new spatiotemporal graph convolutional network (STGCN) for the visual ambiguity of similar negative samples. The author constructs a graph model by connecting all the horizontal bars in different frames to model the temporal relationship. Considering the structural information within the frame, the author constructs a patch map for each frame in the video to provide supplementary information about the appearance. The purpose is to provide complementary information between different patches [6]. The use of such a large number of sample frames will cause more background noise and algorithm cost. This paper considers classifying the video sequence of pedestrians based on the reference pose, and extracts the sample frame with the highest image quality and the most complete spatial information in this reference pose to reduce the influence of occlusion, pose change and redundant information on re-identification. First, estimate the joint points and skeleton information of pedestrian video sequences based on the pedestrian skeleton estimation method proposed by Cao *et al.* The skeleton estimation method uses deep learning neural network and convolution operation to process the image, outputs a heat map for each joint point of the human body, and uses the peak value to indicate the position of the joint point [7]. This paper conducts posture alignment based on the posture alignment module, and the three indicators of distance, angle and confidence between important joint points are measured. These three indicators fully reflect the similarity of posture and image quality. The pose alignment module selects the most discriminative image based on the degree of occlusion of the image and the quality of the image to construct a sample frame set. To reduce the influence of background noise, this paper divides the multi-scale region based on pedestrian joint points, and proposes a posture-guided multi-scale structural relationship learning model.

The study found that the pedestrian re-identification method based on local features has superior performance. Yin *et al.* proposed a local posture stream based on specific joints to extract local dynamic posture features of pedestrians to reduce the impact of noise in the action feature map caused by non-human action on network learning. They divide the human body into head, upper body, arms, legs and other parts based on specific joints, and each part extracts local dynamic features. This modeling method is expected to improve the robustness of local dynamic posture features [8]. Patruno *et al.* proposed a color descriptor based on the Skeleton Standard Pose (SSP). First, the posture adjustment is performed through the skeleton information of the query set to create the skeleton standard posture partition grid, as shown in Fig. 1(c). Then the skeletal standard pose grid is applied to all candidate sets, the color features of the partition grid are extracted, and the color features of the partition grid are weighted to obtain the person color descriptor [9]. Cai *et al.* proposed a multi-scale body mask to guide attention network, which uses body part masks to guide corresponding attention training. They creatively used masks of different parts of the body to guide attention learning, dividing the human body into upper and lower parts, which were respectively used to guide the training of attention on the upper and lower parts. This method can effectively reduce the influence of pedestrian posture changes and background clutter [10]. These methods divide the human body into different components and use separate branches to represent them, but these methods ignore the relationship between different components. Through experiments, this paper found that this constraint relationship plays an important role in the feature learning process, which can make the feature representation of network learning more discriminative. The parts of the human body at different positions can be divided into various regions through a priori knowledge. Through the fusion of the relational model and the graph convolutional network, the local features are constrained by the context information of the feature space, to further refine the learning. Based on the sampled frame images in the reference pose, this paper conducts multi-scale structural relationship learning. Based on the joint points, the pedestrian is divided into three forms, namely the joint point region, the region between the joint points and the body semantic region, as shown in Fig. 1(d)-(f). Based on the contextual constraint relationship between regional features, this paper constructs a relationship model, and transfers the relationship scores learned by

the relationship model to the node propagation layer of the graph convolutional network. We process and normalize the relationship scores as the relationship weights of node propagation, and use the graph convolution network based on the relationship weights to extract structural relationship features. The multi-scale regional structure relationship feature enhances the discrimination ability of local features in the pedestrian re-identification problem. The main contributions of this paper are as follows,

(i) A posture-guided pedestrian re-identification framework is proposed to optimize the learning ability of local information in the part-based method and enhance the distinguishing ability of local features in pedestrian re-recognition problems.

(ii) A method of multi-scale region division of the human body is proposed. The relationship between the local features of pedestrians is constrained by constructing a relationship model, and the local feature pairs are used as input conditions to solve the relationship matrix between the local features.

(iii) The relationship weight is derived based on the relationship matrix, and the transfer learning method is adopted to transfer the relationship weight to the node propagation layer of the GCN to learn the structural relationship feature of the pedestrian multi-scale region.

II. RELATED WORK

A. PART-BASED RE-ID

Due to the lack of fine-grained detailed representation of global features, local-based methods have attracted more and more attention from researchers, which is consistent with the recognition mechanism of the human brain. When recognizing objects or people, their first reaction is to look for significant local discriminant regions, and the recognition effect is better based on multiple discriminant regions and global information. Researchers divide the feature map of the input image equally in the vertical direction, and generate a partial feature representation of the person through a predefined fixed height [11]–[13]. In order to solve the occlusion problem, Yang *et al.* proposed a novel spatiotemporal graph convolution network (STGCN) to model the temporal relationship between different frames and the spatial relationship within the frame. For each frame of the image, the equally divided horizontal feature map is used to construct the structure graph convolution model, and the temporal graph convolution model is constructed based on the horizontal feature map of all frames of the video sequence to capture the temporal dynamic relationship between different frames [6]. For the problem of people's misalignment in the image, Sun *et al.* proposed a partial convolutional baseline (PCB) to learn discriminative partition features. The feature map is divided into 6 equal parts by the PCB network, and average pooling and dimensionality reduction are performed. For the regional outliers generated by this processing method, the refined part pooling (RPP)network is further proposed, and these outliers are re-allocated to the closest region to generate accurate local features with internal consistency [14]. Yao *et al.* used the region of interest pooling method to estimate various parts of the human body in the feature space by detecting local regions of the human body, and aggregated the local features of person [15]. Yang *et al.* used the posture estimation model to detect the joint points of person. They divided the person into 5 different parts according to the human body structure, and extracted the features of each part for re-identification [16]. Jing *et al.* proposed a posture-guided joint point global and local matching network, introduced posture as supervision information, and combined noun description features and visual features to obtain posture attention feature maps [17].

B. VIDEO-BASED RE-ID

Video person re-identification is a further expansion of pedestrian re-identification based on images, because the video contains more spatial information, and most importantly, the video sequence contains temporal cues of person. For occlusion problems, video person re-identification can supplement the features of the occluded part based on the temporal relationship between frames, which can further reduce the influence of factors such as viewing angle, occlusion, and posture changes.

How to extract discriminative video-level feature representations is a key issue in video person re-identification. In recent years, deep learning technology has achieved remarkable success in the field of computer vision. Therefore, the person re-identification method based on deep learning has been favored by more and more researchers. Song *et al.* used CNN to extract spatial features while using Recurrent Neural Network (RNN) to extract temporal features. Aiming at the problem of insufficient image information in a single frame, they considered complementing each other for image information of multiple frame sequences. Song *et al.* evaluate the quality of the picture region, and then compensate the high-quality regions from other frames to the low-quality regions [18]. Traditional Long Short-Term Memory (LSTM) network uses the hidden layer between the upper and lower frames to learn feature representations with temporal information. Liu *et al.* proposed Refining Recurrent Unit (RRU) to upgrade inter-frame features. Unlike LSTM, which learns new features from temporal feature vectors, RRU does not directly use the features of each frame to extract temporal information, and restore the missing parts of the current frame according to the appearance and context of historical video frames [19]. Liu *et al.* proposed an end-to-end Comparative Attention Networks (CAN) based on soft attention. The end-to-end model can selectively focus on the distinguished parts of the image after learning several person pictures, use comparative attention components to generate attention regions, and generate attention maps through LSTM. The CAN model can simulate the human perception process to verify whether the two images are the same person [20]. Wu *et al.* proposed a global deep video representation learning method as a supplement to the 3D convolutional neural network layer to capture the appearance and motion dynamics in the

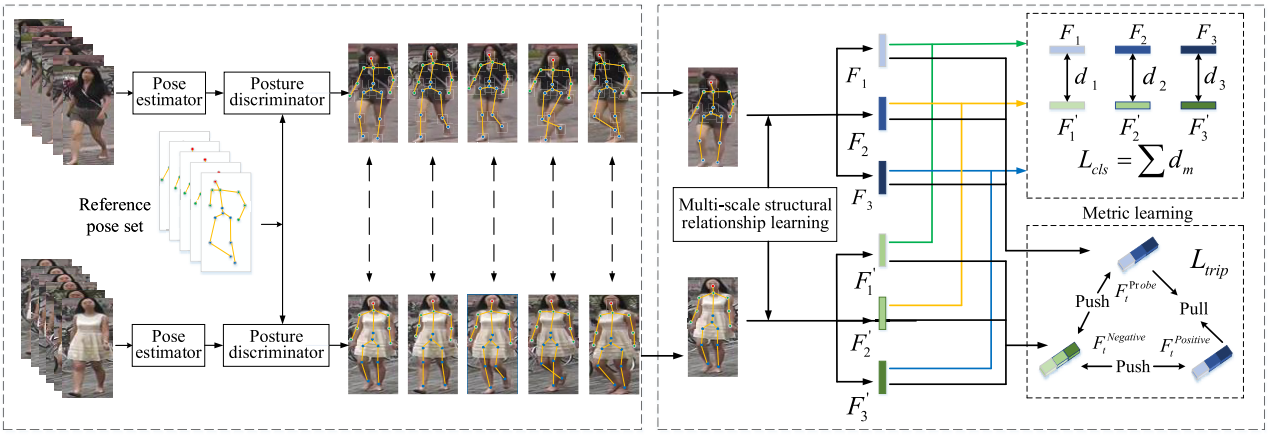


FIGURE 2. Framework flow chart. (a) Posture alignment module; (b) Multi-scale structural relationship learning and metric learning module.

video, and aggregate local three-dimensional features across the entire video. The proposed network further increases the 3D local alignment method and learns local features in a soft attention method [21]. To solve the problem of image quality output fluctuations over time, Chen *et al.* proposed a Spatio-temporal Attention-aware Learning (STAL) model based on video person re-identification. It aims to focus on the important part of person in the video in Spatio-temporal domains, which helps to extract and adaptively match person features [22]. To solve the problem of occlusion and visual ambiguity of visually similar negative samples, Yang *et al.* modeled the temporal relationship between different frames and the spatial relationship within frames, and proposed a new spatiotemporal graph convolution network (STGCN). STGCN includes two GCN branches, the spatial branch extracts the structural information of the human body, and the temporal branch extracts discriminative information from adjacent frames [6]. In order to make full use of the relationship between the various parts, Wu *et al.* proposed a novel adaptive graph representation learning method, which realizes the contextual interaction between the features of related regions. An adjacency graph with adaptive structure perception is constructed through posture alignment connection and feature affinity connection, and the internal relationship between graph nodes is modeled [5].

C. ATTENTION MECHANISM IN RE-ID

Due to the limitations of the target detection algorithm, the bounding box detected by pedestrians is not accurate enough. For example, only part of the pedestrian contour in the image can be detected. This unconstrained automatic detection error has a great impact on pedestrian re-identification. In order to solve this problem, researchers proposed an attention mechanism to help the network learn important information about pedestrians. The attention mechanism is playing an increasingly important role in the field of computer vision. Li *et al.* introduced a coordinated attention method and constructed a new harmonious attention convolutional neural network (HA-CNN) model for joint learning of soft pixel attention and hard region attention, while optimizing feature representation to solve misaligned images of person re-identification problems [23]. Liu proposed a multi-direction attention model, which uses attention mapping to mask features at different levels and further extract attention features [24]. At present, most attention-based methods generate global attention through the entire image, ignoring the local attention of body parts, and perform poorly in the face of posture changes, misalignment, and partial occlusion. Cai *et al.* proposed a Multi-scale Body-part Mask Guided Attention network (MMGA), which uses body-part masks to guide the training of corresponding attention, which can simultaneously learn global and local attention to help extract global and local features [10]. Most methods prefer rough first-order attention, such as spatial and channel attention, and rarely explore higher-order attention mechanisms. Chen *et al.* proposed a hybrid high-order attention network to model and used complex and high-order statistical information to capture the subtle differences between pedestrians, thus generating discriminative attention suggestions [25].

III. METHOD

In this section, we introduce the proposed video person re-identification method for pose-guided multi-scale structural relationship learning(PMSRL). The method mainly includes pose alignment module, multi-scale structural relationship learning and metric learning module. The flow chart of the entire framework is shown in Fig. 2.

A. POSTURE ALIGNMENT BASED ON REFERENCE POSE

For the processing of redundant information in the video sequence, we consider the selection of keyframes based on the reference pose. This paper defines the distance metric,

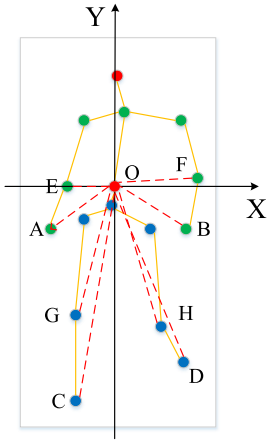


FIGURE 3. Schematic diagram of distance metric and angle metric between the joint points of person posture.

angle metric and confidence degree between joint points as the basis for selecting keyframes. In the video sequence, pedestrians will present images of different angles and postures. In these images, the positions of the pedestrian's head, shoulders, and crotch are relatively fixed, and the joint points at the end of limbs vary greatly. Therefore, the position information of the joint points at the end of the limb is the most ideal representation of posture. In order to better characterize the person posture, we calculate the metrics based on the position information and confidence of the joint points at the end of the person limbs. As shown in Fig. 3, taking the pedestrian's center of gravity O as the origin, the horizontal direction to the left as the positive direction of the X axis, and the vertical direction upward as the positive direction of the Y axis. The distance and angle of the end joint point relative to the center of gravity O are measured based on the pedestrian coordinate system. The metric index is,

$$I = \{D_{Oi}, \theta_{Oi}, C_i | \, i \in A, B, C, D, E, F, G, H\}, \tag{1}$$

$$\begin{cases} D_{Oi} = [d_{OA}, d_{OB}, d_{OC}, d_{OD}, d_{OE}, d_{OF}, d_{OG}, d_{OH}] \\ \theta_{Oi} = [\theta_{OA}, \theta_{OB}, \theta_{OC}, \theta_{OD}, \theta_{OE}, \theta_{OF}, \theta_{OG}, \theta_{OH}] \\ C = \sum C_i, i \in A, B, C, D, E, F, G, H, \end{cases} \tag{2}$$

D_{Oi} is the matrix representation of the Mahalanobis distance between the joint point and the coordinate circle point O, θ_{Oi} is the matrix representation of the angle measurement between the joint point and the coordinate circle point O, C_i is the joint point confidence, C is the sum of the joint point confidence, used to reflect the degree of occlusion of the image and the image quality, As the final evaluation index, I is a collective expression form of the distance, angle, and confidence metric factors of pedestrian limbs.

The Mahalanobis distance between the joint point and the coordinate circle point O is calculated by the coordinate position,

$$d_{Oi} = d_M^2(x_O, x_i) = (x_O - x_i)^{\mathrm{T}} M (x_O - x_i). \tag{3}$$

The angle metric between the joint point and the coordinate circle O is expressed by the arcsine function of the distance,

$$\theta_{Oi} = \arcsin\left(\frac{x_i}{d_{Oi}}\right) = \arcsin\left(\frac{x_i}{(x_O - x_i)^{\mathrm{T}} M (x_O - x_i)}\right). \tag{4}$$

The degree of occlusion and quality of the image is calculated by the joint point confidence, which is obtained by the pedestrian skeleton estimation method of Cao *et al.* [7],

$$C = \sum C_i, i \in A, B, C, D, E, F, G, H. \tag{5}$$

Based on the reference pose set P_n, the expression results of the corresponding pose evaluation index I in the video sequence are respectively calculated. By comparing the results of the evaluation index, the sampling frame with the most complete information and the highest image quality under the pose is selected to form the sampling frame set $\{P_t, t = 1, 2 \cdots T\}$, which can solve the problem of pedestrian re-identification more efficiently.

B. MULTI-SCALE STRUCTURAL RELATIONSHIP LEARNING

1) RELATIONAL MODEL

At present, most of the local-based pedestrian re-identification methods divide the human body into head, shoulders, upper body, arms, and legs. Different parts are often represented by a single branch. However, this method ignores the relationship between different parts. This relationship can play a restrictive role in the feature learning process, and can directly make the result of feature learning more robust. The parts of the human body at different positions can be divided into multiple scales through a priori knowledge, and the local feature context information can be constrained through the relational modeling method, to further refine the learning.

As shown in Fig. 4, the main idea of the relationship model is to learn the relationship matrix between local features through the context constraints of local features. The relationship matrix is calculated in the form of feature pairs, and $K \times K$ feature pairs are constructed based on the local features of pedestrians $K \times C$, where K is the number of local features, and C is the dimension of local features. The feature pair is mapped to the feature space of C_r dimension through multi-layer perception, and the relationship mapping tensor obtained is,

$$r_{i,j}^T = p\left(h_i, h_j\right) + p\left(h_j, h_i\right), \tag{6}$$

$r_{i,j}^T$ represents the relational mapping tensor of $K \times K \times C_r$, using 1×1 convolution to perform dimensionality reduction operations, and the output relation matrix channel dimension is 1, each element $a_{i,j}$ represents the relational weight between the i^{th} local feature and the j^{th} local feature. The relation matrix is expressed as,

$$r_{i,j} = \sigma\left(r_{i,j}^T\right), \tag{7}$$

σ represents 1×1 convolution, BN, ReLU operation.

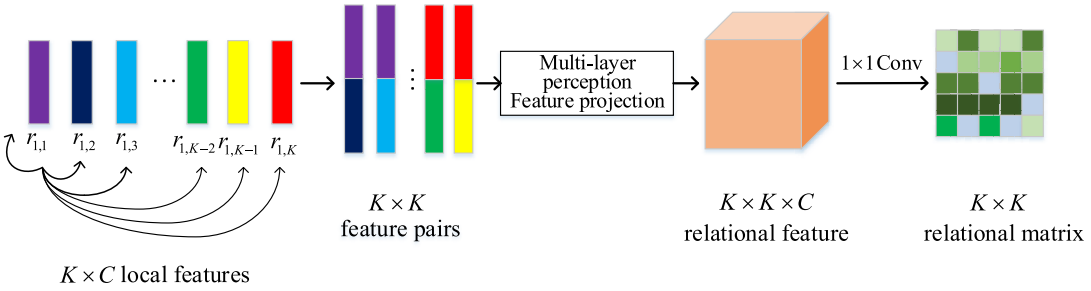


FIGURE 4. Flow chart of relationship model.

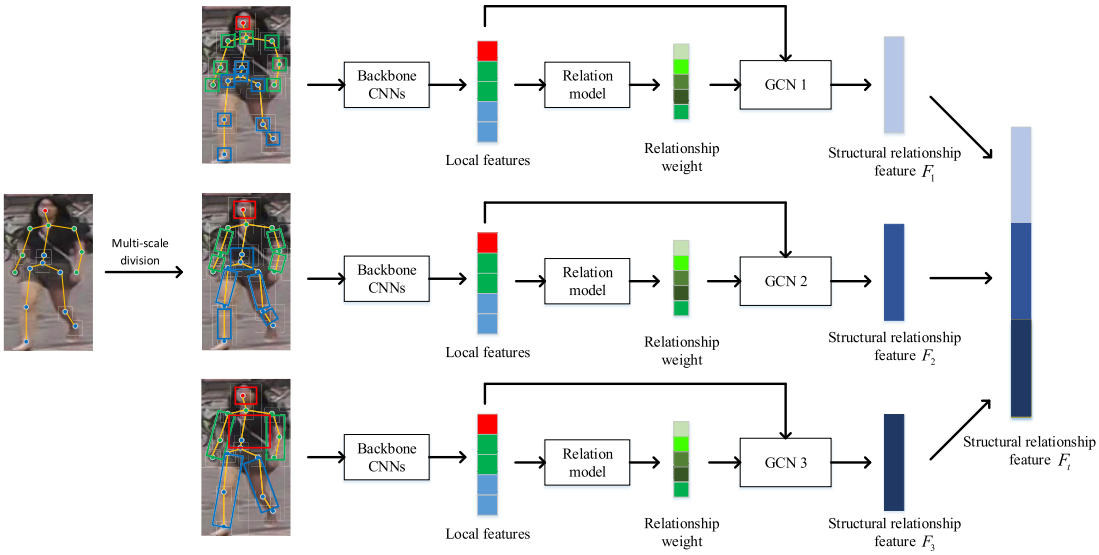


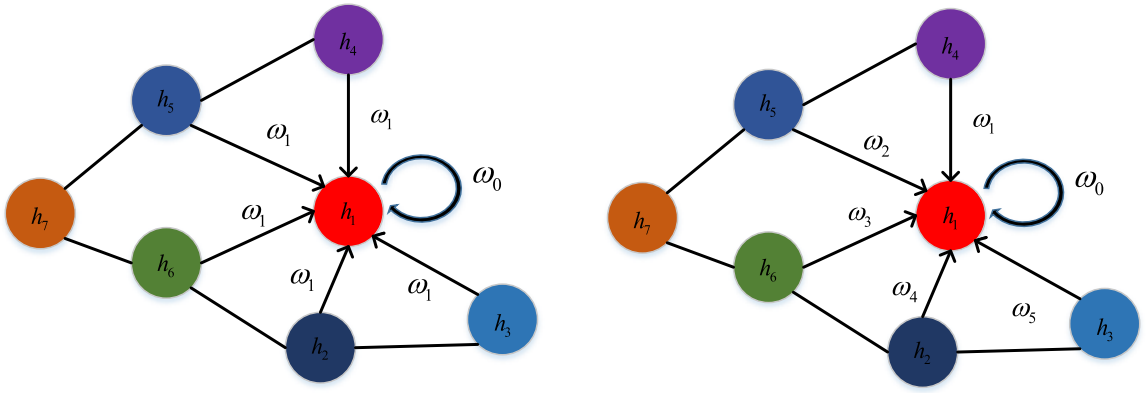
FIGURE 5. Flow chart of multi-scale structural relationship feature extraction.

2) MULTI-SCALE STRUCTURAL RELATIONSHIP LEARNING

In addition to the differences in a walking posture, there are also differences in clothing among pedestrians. In order to realize the re-identification of pedestrians with different wearing styles, this paper considers the multi-scale division of pedestrians, and divides pedestrian into joint point region, the region between joint points and the body semantic region. The joint point region is centered on the joint point, and a square region is obtained by expanding with a certain width α. The joint point region can better retain the overall posture and movement information of the pedestrian. The region between the joint points is based on two joint points as the two ends of the region, and a rectangular region is obtained by expanding with a certain width β. The region between the joint points has a stronger discriminability for the difference of wearing of pedestrians (such as long sleeves and short sleeves, trousers and shorts). Body semantic region is to divide the human body into head, upper body, left arm, right arm, left leg, right leg and other regions according to the semantic description. Body semantic region can reduce occlusion and enhance the ability to distinguish left and right information of pedestrians. The idea of multi-scale region division is the expression of multi-level information of pedestrians, the flow chart of multi-scale structure relationship feature extraction is shown in Fig. 5. The three branches of the flowchart corresponding to the above three types of local regions. Compared with the horizontal striping and dense grid method, the feature extraction method in this paper is based on the joint points to locate the local region, which is more generalized for scenes such as viewing angle change, complex background, and wearing.

The feature extraction of traditional images is based on convolutional neural networks, but the skeleton of pedestrians is a non-European shape, so convolutional neural networks are not applicable. This paper considers the use of graph convolutional network (GCN) for pedestrian feature extraction. The advantage of graph convolutional network lies in the feature transmission method between nodes, and the structural relationship between pedestrian local regions can be retained and learned. The node weights of the traditional graph convolutional network are shown in Fig. 6(a), which includes a self-connected weight ω_0, and weight sharing among other nodes,

$$\omega_1 = \frac{1 - \omega_0}{n} \tag{8}$$



(a) Traditional GCN node weights

(b) GCN node relationship weight

FIGURE 6. Comparison of GCN node weights.

In this paper, the transfer learning method is used to transfer the relationship matrix learned by the relationship model to the node propagation layer of the GCN, and the relationship matrix is reprocessed based on the adjacency matrix of the pedestrian region to obtain the relationship weight of the node propagation. The expression form of the graph structure between local regions is reflected by the adjacency relationship between nodes, the adjacency matrix A,

$$A_{i,j} = \begin{cases} 1 & h_i \text{ is adjacent to } h_j \\ 0 & otherwise \end{cases} \tag{9}$$

here h_i is the i^{th}local feature of the pedestrian, which represents the i^{th} node feature. If h_i and h_j are adjacent to each other, then $A_{i,j} = 1$, otherwise $A_{i,j} = 0$, and the adjacency matrix A is symmetric. The relationship matrix contains the relationship scores of all feature pairs, and the relationship scores of non-adjacent feature pairs are redundant for the node propagation layer of the GCN. Therefore, this paper uses the adjacency matrix to reprocess the relationship matrix to obtain the relationship weight of the node propagation. The normalized expression of the relationship weight is,

$$\omega_{i,j} = \varepsilon \left(A \otimes r_{i,j} \right) \tag{10}$$

$\otimes$ is the Hadamard product, which represents the product of the corresponding positions of the matrix, and $\varepsilon(\cdot)$ represents the normalization operation. In GCN, each hidden layer maps high-dimensional node features to low-dimensional feature space, and then spreads node information to each other. In this paper, a three-layer GCN is used to extract the structural relationship feature between local regions, and the expression of node feature propagation is,

$$h_i^{(l+1)} = \sigma \left(\omega_0^{(l)} h_i^{(l)} + \sum_{j \in N_i} \frac{1}{c_{ij}} \omega_{i,j}^{(l)} h_j^{(l)} \right) \tag{11}$$

here N_i represents the set of adjacent nodes of node i, and c_{ij} is the symmetric normalization constant of the adjacency matrix. $\sigma(\cdot)$ represents the nonlinear activation function ReLU. In order not to lose the feature information of the node, each node will have a self-connection (as shown in Fig. 6), which is equivalent to adding an identity matrix I to the adjacency matrix A, $\widetilde{A} = A + I_N$. Symmetric normalization of adjacency matrix,

$$\hat{A} = \widetilde{D}^{-\frac{1}{2}} \widetilde{A} \widetilde{D}^{-\frac{1}{2}}. \tag{12}$$

The structural relationship feature of the output of the three-layer GCN are,

$$\begin{aligned} F_m &= f\left(H^{(0)}, A\right) \\ &= \hat{A}\left(\hat{A}\, \mathrm{ReLU}\left(\hat{A} H^{(0)} W^{(0)}\right) W^{(1)}\right) W^{(2)}, \end{aligned} \tag{13}$$

here $H^{(\cdot)}$ represents the feature representation of all nodes of the layer, $W^{(0)} \in R^{N \times D_0}$ is the weight matrix of the relationship between the input layer and the hidden layer, $W^{(1)} \in R^{N \times D_1}$ is the weight matrix of the relationship between the hidden layer and the hidden layer, and $W^{(2)} \in R^{N \times D_2}$ is the weight matrix of the relationship between the hidden layer and the output layer. D_0, D_1, and D_2 respectively represent the dimensions of the input layer, hidden layer, and output layer. The multi-scale structural relationship feature based on a single sample frame is expressed as,

$$\begin{aligned} F_t &= \sum_{m=1}^{3} F_m = \sum_{m=1}^{3} f\left(H^{(0)}, A\right) \\ &= \hat{A}\left(\hat{A}\, \mathrm{ReLU}\left(\hat{A} H^{(0)} W^{(0)}\right) W^{(1)}\right) W^{(2)} \end{aligned} \tag{14}$$

The video-level features based on the reference pose sets are expressed as,

$$F_T = \sum_{t=1}^{T} F_t = \sum_{t=1}^{T} \sum_{m=1}^{3} F_m. \tag{15}$$

C. METRIC LEARNING

The region division of the pedestrian image is based on the joint points. For the measurement based on the local region, this paper uses the structural relationship features of the corresponding region to metric the distance for the three types of local regions in each sampling frame, as shown in the metric learning module in Fig. 2(b). The distance metric loss is used to express the difference of the appearance color of the pedestrian multi-scale regions. For the structural relationship features of multi-scale regions, the distance metric loss expression is,

$$L_{cls} = \sum_{m=1}^{3} d_m = \sum_{m=1}^{3} \left\| F_m - F'_m \right\|_M^2 = \sum_{m=1}^{3} \left(F_m - F'_m \right)^{\mathrm{T}} M \left(F_m - F'_m \right) \quad (16)$$

here, F_m represents the structural relationship feature of the multi-scale region of the pedestrian to be detected, and F'_m is the structural relationship feature of the multi-scale region corresponding to the pose of the pedestrian to be detected in the candidate set.

This paper uses triple loss for deep metric learning. The main idea is that in a batch, the triples are composed of the sample to be tested, a positive sample and a negative sample. Randomly sample Q pedestrians (pedestrian identities), each pedestrian obtains T sampling frames based on the reference pose, so a batch of QT images are generated, and the triplet loss is,

$$L_{triplet} = \sum_{t=1}^{T} [\max D(F_t^{\mathrm{Pr}obe}, F_t^{Postive}) - \min D(F_t^{\mathrm{Pr}obe}, F_t^{Negative}) + a]_+ \quad (17)$$

a is the set threshold parameter, the final total loss is equal to the sum of the two loss functions,

$$L = L_{cls} + L_{triplet}. \quad (18)$$

IV. EXPERIMENT

In this section, the method proposed in this paper will be verified on three challenging video data sets, including iLIDS-VID [26], PRID 2011 [27], MARS [28]. First, the data set and experimental settings are introduced in Section 3.1, the components and parameters of the proposed method are evaluated in Section 3.2, and the comparison with the latest method is in Section 3.3.

A. DATASETS AND SETTINGS

1) iLIDS-VID

The dataset involved in a public open space in the two disjoint camera view observed 300 different pedestrians, in monitoring aircraft in the docking station hall consists of two disjoint the video camera to create the data set. The data set randomly sampled 600 videos of 300 people, each with two video clips, and the average length of the video clip is 73 frames. The dataset was challenging for similar clothing, visual change, complex background, and severe occlusion.

2) PRID 2011

This dataset provides the trajectory of multiple pedestrians under two static surveillance cameras. One camera shoots 385 people, the other shoots 749 people, and 200 people at the same time appear in two perspectives. Each video has 5 frames to 675 frames, the average video length is 100 frames. The dataset was collected in an uncongested outdoor scene with a relatively simple and clean background and less occlusion.

3) MARS

This dataset is currently the largest video pedestrian re-identification data set, which is expanded from the market1501 data set. It is composed of videos shot by 6 cameras at the same time, including 17,503 real video sequences and 3,248 interference video sequences, a total of 20751 video sequences of 1261 different pedestrians. The data set is divided into a training set of 625 people and a test set of 636 people. Each person has an average of 13 video sequences. Everyone has at least two videos taken by different cameras. These video sequences are generated by DPM detectors and GMMCP trackers, making the identification of this data set more challenging.

These three databases contain the problems of illumination, perspective, posture change and occlusion faced by pedestrian re-recognition. All experiments are repeated 10 times and the average accuracy is calculated to reduce random errors. The performance of the algorithm is mainly measured according to the cumulative matching curve (CMC) and mAP. The CMC curve represents the probability that the algorithm finds the correct match among the top-k results with the highest similarity. In all experiments, pedestrian images are processed as 512×256 pixels.

In the experiment, the pedestrian skeleton estimation model proposed by Cao *et al.* is used to extract the joint points and skeleton information of pedestrian images, which provides the position and confidence of joint points for the selection of key frames of video sequences. GCN is first pre-trained on the ImageNet dataset, and then fine-tuned on the iLIDS-VID, PRID 2011, MARS dataset. In the training phase, 50% of the data set samples are randomly selected as training samples, and the stochastic gradient descent algorithm (SGD) is used to update the network learning. The initial learning rate is set to 0.1 and gradually close to 0.01. In the comparative experiment of reference pose and random pose, this article sets the number of poses as $T = 8$, and the size of the multi-scale division is fine-tuned in subsequent experiments. In the test phase, the experiment is repeated 10 times to calculate the average accuracy of Rank-1, Rank-5, and Rank-20 as the model's pedestrian re-identification performance evaluation index.

B. EVALUATION OF COMPONENTS AND PARAMETERS

In this section, we will evaluate the effectiveness of all components of the proposed method and the setting of parameters, including the number of poses, multi-scale region size parameters, occlusion analysis, and ablation experiments.

1) ANALYSIS OF THE NUMBER OF POSTURES AND THE IMPORTANCE OF REFERENCE POSTURES

Reasonable and efficient use of video information is an important indicator of algorithm performance. Considering the cost of the algorithm, this paper selects sampling frames based on a set of important reference poses. In order to highlight the performance of the reference posture, this paper conducts comparative experiments to verify the randomly selected posture group. A group of randomly selected postures is called a random posture group. The number of postures is a good representation of video information. If the number of postures selected is too small, the video sequence information will not be fully utilized. If the number of selected postures is too large, it will cause posture confusion, consume more computing time, and the performance of the algorithm will show a tendency to decrease. Therefore, this paper conducts an experimental analysis on the number of postures and the importance of reference postures based on the PRID 2011 data set.

In this paper, a sampling frame is selected for experimental analysis under each pose. Fig. 7 shows that when $T < 8$, as the number of postures increases, the accuracy of Rank-1 and Rank-5 continue to rise, reaching the maximum at $T = 8$. At this time, the accuracy of Rank-1 reaches 94.7%, the accuracy rate of Rank-5 reaches 99.2%. After that, the recognition accuracy gradually decreases as the number of sampled frames increases, and gradually approaches the fitting state. This phenomenon is caused by pose confusion. Moreover, a large number of pose sampled frames causes more resource consumption and additional costs. Fig. 8 is the comparison result between the random posture group and the reference posture group. To avoid errors caused by random phenomena, this article sets up multiple sets of random postures for experimental analysis. Fig. 8 only lists some of the results. It can be seen from the experimental results that Rank-1, Rank-5, Rank-10, and Rank-20 of the reference pose group are higher than those of the random pose group. Experiments show that the pedestrian re-identification method based on posture guidance has good performance in solving the problem of video information redundancy.

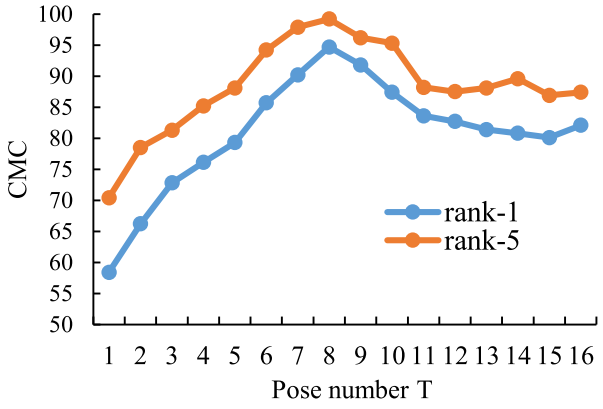


FIGURE 7. Experimental analysis of the number of postures.

2) SIZE ANALYSIS OF MULTI-SCALE REGIONS

To fully reflect the effectiveness of the local feature pedestrian re-identification method, this paper divides the sampled frame into multiple local regions. Under different postures and perspectives, the joint point region, the region between joint points, and the limb semantic region all show superior performance. For the frontal image of a pedestrian, the region between the joint points such as the head, arms, and legs is more discriminative. For the image of the side of the pedestrian, the joint point region and the semantic region of the limbs are more discriminative. This section analyzes the influence of the size of the local region on the result of re-identification through experiments. For the region between the joint points and the semantic region of the limbs, the joint point position is used as the reference condition, and the learning of the diversity regularization constraint region is adopted. The visualization result is shown in Fig. 9. In different postures and perspectives, the size of the region between the joint points and the semantic region of the limbs is different. To further stabilize the performance of the algorithm, this section accurately evaluates the size of the joint point region. The size of the joint point region is set to eight sizes of 10×10, 11×11, 12×12, 13×13, 14×14, 15×15, 16×16, 17×17, and the accuracy of Rank-1 is recorded on the iLIDS-VID and PRID 2011 data sets. The experimental results are shown in Fig. 10.

Fig. 9 shows that the size of the joint point region has a certain degree of influence on the accuracy of re-identification. From the peak of the curve, it can be concluded that on the PRID 2011 data set, when the size of the joint area is 14×14, the accuracy of Rank-1 reaches 94.7%. On the iLIDS-VID data set, when the joint region size is 14×14, the accuracy of Rank-1 reaches 85.5%. The reason why the curve exhibits such fluctuations is that when the size of the joint point region is too large, additional background noise will be added, and it will also cause confusion between the joint point regions, resulting in a gradual decrease in the recognition effect. When the joint point region is too small, the discriminative ability of the regional feature is weak, which ultimately leads to poor re-identification accuracy. Through the experimental verification of two data sets, the optimal value of the joint point region size in this paper is set 14×14.

3) OCCLUSION ANALYSIS

In data sets and real scenes, occlusion is an unavoidable factor. According to the occlusion situation that occurs in the real scene, this section focuses on the left and right occlusion of pedestrians. The probability of occlusion of the upper

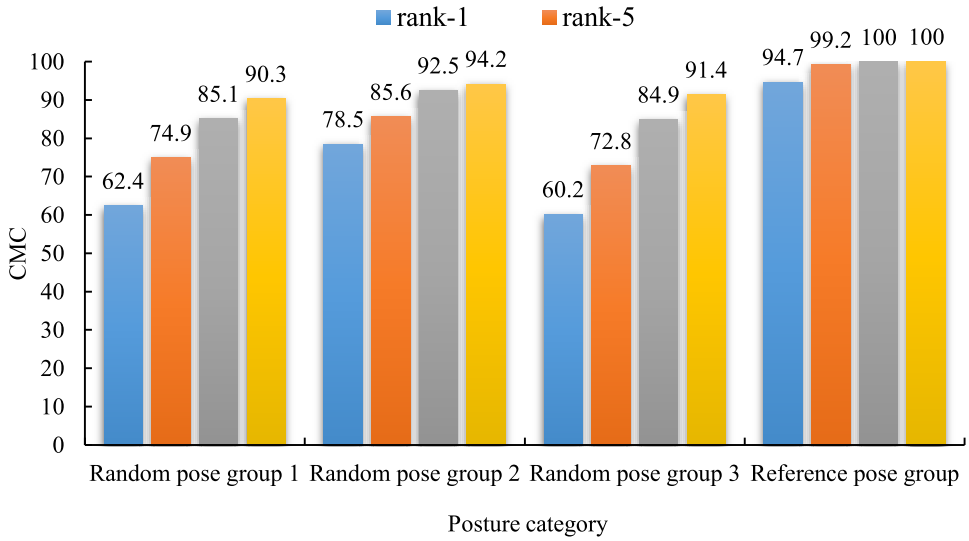


FIGURE 8. Comparison of results between random posture group and reference posture group.

FIGURE 9. Visualization results of the region between the joint points and the semantic region of the limbs.

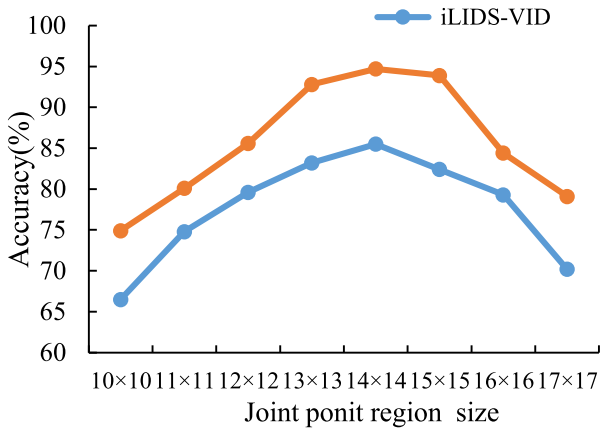


FIGURE 10. Rank-1 accuracy curve of the joint point region size of the iLIDS-VID and PRID 2011 datasets.

and lower body of pedestrians is relatively low. Therefore, corresponding to the three region division methods in this paper, corresponding occlusion methods are set respectively. The occlusion category I am the occlusion of a single joint point region (joint point region occlusion), the occlusion category II is the occlusion of the adjacent joint point region (region occlusion between joint points), and the occlusion category III is the occlusion of the entire arm or leg (occlusion of the semantic region of the limbs). According to the occlusion category and region division method, the corresponding occlusion experiments are carried out. The optimal value of the above experiment was selected for the number of postures and the size of the joint region. The results of the occlusion experiment are shown in Tab. 1.

In Tab. 1, as the degree of occlusion increases, the accuracy of pedestrian re-identification gradually decreases. First, each joint point region is individually occluded to obtain the accuracy rate of a single joint point occlusion, and then the average of these 16 accuracy rates is calculated to obtain the accuracy rate of occlusion category I. From the data with no occlusion, occlusion category I and occlusion category II, it can be found that the occlusion of a single joint point and a small region has little effect on the recognition result. At the occlusion category III, the accuracy of Rank-1 of our method on the iLIDS-VID dataset can still reach 65.4%, which shows that our method has a good performance in the problem of occlusion.

C. COMPARISON WITH THE STATE-OF-THE-ART METHOD

This paper further evaluates the performance of the PMSRLRe-id method and compares it with the matching results of the most advanced methods on iLIDS-VID, PRID 2011 and MARS data sets. To better highlight the robustness of the proposed method to multi-scale regions, we set a simple

TABLE 1. Recognition results under different occlusion categories.

Occlusion	iLIDS-VID		PRID 2011	
category	Rank-1	Rank-5	Rank-1	Rank-5
None	85.5	91.4	94.7	99.2
I	85.3	91.4	94.6	99.2
II	78.6	84.1	88.4	94.4
III	65.4	72.7	76.8	84.6

TABLE 2. Comparison results of different methods in ilids-vid dataset.

Dataset	iLIDS-VID			
Methods	Rank-1	Rank-5	Rank-10	Rank-20
SPW[29]	69.3	89.6	95.7	98.2
DGM [30]	42.6	67.7	76.6	85.8
SI^2DL[31]	48.7	81.1	89.2	97.3
SPRNN[32]	55.2	86.5	--	97.0
ASTPN[33]	62.0	86.0	94.0	98.0
QAN[4]	68.0	86.8	95.4	97.4
FARL[34]	68.4	87.2	--	--
M3D[35]	74.0	94.3	--	--
AGRL[5]	84.5	**96.7**	--	99.5
PISA[36]	85.4	96.7	92.5	99.5
Snippet[37]	85.4	96.7	--	99.5
AMOC [38]	68.7	94.3	98.3	99.3
Baseline	72.3	87.5	96.6	99.5
Ours	**85.5**	91.4	**99.3**	**100**

baseline, which only considers the structural relationship feature of the joint point region. The experimental results are shown in Tab. 2, Tab. 3, and Tab. 4.

1) RESULTS ON iLIDS-VID

As shown in Tab. 2, on the iLIDS-VID data set, we conduct ablation experiments. When using the joint point regional structural relationship features alone, the accuracy of Rank-1 is only 72.3%. It is because the regional features of the joint points only consider the drawbacks brought about by the regional features related to movement changes, and a large amount of other spatial information is ignored. In the PMSRLRe-id method, the performance of multi-scale regional features compensates for each other. The accuracy rate of Rank-1 reaches 85.5%, and the accuracy rate of Rank-10 reaches 99.3%, which is better than the accuracy of existing methods. The Rank-1 of our proposed method is 13.2% higher than the baseline, which verifies the importance of multi-scale local regions for re-identification.

TABLE 3. Comparison results of different methods in PRID-2011 dataset.

Dataset	PRID-2011			
Methods	Rank-1	Rank-5	Rank-10	Rank-20
SPW[29]	83.5	96.3	98.5	100
DGM [30]	83.3	96.7	98.9	99.6
SI^2DL[31]	80.3	96.5	97.9	99.5
SPRNN[32]	79.4	94.4	--	99.3
ASTPN[33]	77.0	95.0	99.0	99.0
QAN[4]	90.3	98.2	99.3	100
FARL[34]	91.2	98.9	--	--
M3D[35]	94.4	**100**	--	--
PISA[36]	92.9	98.3	99.1	99.9
AGRL[5]	94.6	99.1	--	100
Snippet[37]	93.0	99.3	--	100
AMOC [38]	83.7	98.3	99.4	100
Baseline	84.2	93.8	99.4	99.9
Ours	**94.7**	99.2	**100**	**100**

TABLE 4. Comparison results of different methods in MARS dataset.

Dataset	MARS			
Methods	Rank-1	Rank-5	Rank-20	mAP
MQ[28]	68.3	82.6	89.4	49.3
M3D[35]	84.4	93.4	97.8	--
QAN[4]	73.3	84.9	91.6	51.7
STMP[19]	84.4	93.2	96.3	72.7
SRL[39]	84.4	93.5	96.8	75.8
AGRL[5]	89.8	96.1	97.6	81.1
STAL[22]	82.8	92.8	98.0	73.5
STGCN[6]	90.0	96.4	98.3	83.7
AMOC [38]	68.3	81.4	90.6	52.9
Baseline	72.6	84.6	91.5	61.6
Ours	**90.2**	**96.6**	**99.8**	**83.2**

2) RESULTS ON PRID 2011

As shown in Tab. 3, on the PRID 2011 data set, our method has achieved the current optimal value of recognition accuracy. Among the methods of video re-identification in recent years, the accuracy of Rank-1 of the pedestrian re-identification method in literature [35] reaches 94.4%, and our method has improved the accuracy of Rank-1 by 0.3% compared with the literature [35]. The accuracy of Rank-10 of the pedestrian re-recognition method in the literature [38] reaches 99.4%, and the accuracy of the Rank-10 method in this paper is 0.6% higher than that of the literature [38].

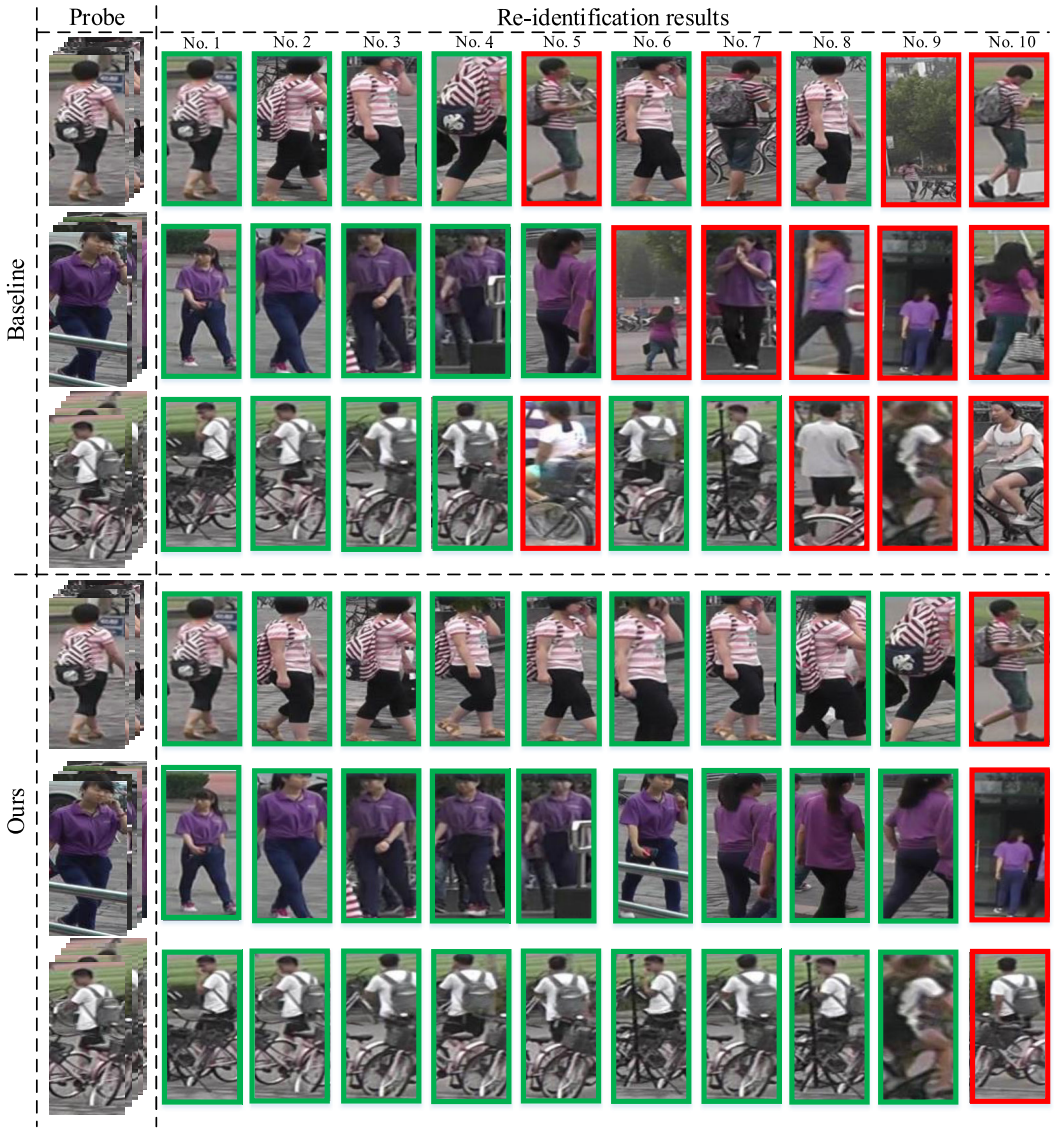


FIGURE 11. Comparison of Rank-10 ranking between the method in this paper and the baseline model.

Compared with the baseline, the accuracy of Rank-1 of the method in this paper increases 10.5%, which further validates the importance of the relationship model between regions.

3) RESULTS ON MARS

As shown in Tab. 4, on the MARS data set, the Rank-1 accuracy rate of PMSRLRe-id method reaches 90.2%, and the accuracy rate of Rank-5 reaches 96.6%. Both indicators are slightly improved compared to STGCN. This method also achieves a significant increase in mAP, reaching 83.2%, which is 2.1% higher than AGRL and 7.4% higher than SRL. Aiming at the problem of visual ambiguity between samples, STGCN models the temporal relationship between different frames and the spatial relationship within the frame. For each frame of the image, the equally divided horizontal feature map is used to construct the structure graph convolution model. Construct a graph map convolution model based on the horizontal feature mapping of the sampled frames of the video sequence to capture the time dynamic relationship between different frames. AGRL proposes an adaptive graph representation learning method based on the contextual interaction between regional features. The image is divided horizontally by joint point positioning to achieve pedestrian posture alignment. This method uses posture alignment connection and features affinity connection to construct an adaptive structural perception adjacency graph. The method proposed in this paper and the above-mentioned methods are based on the local way to realize the interaction between features. The difference lies in the multi-scale division based on the joint points in this paper. The multi-scale division method expresses the pedestrian information at multiple

levels, reducing the influence of posture, background and other factors. The experimental results show that the method in this paper can accurately achieve the re-identification effect and has good generalization.

4) RE-IDENTIFICATION RESULTS

As shown in Fig. 11, we provide the re-identification results of the baseline model and the method in this paper on the MARS dataset. The image with the green box is the positive sample of the probe, and the image with the red box is the negative sample of the probe. The effectiveness of the method in this paper is highlighted through the comparison of the visualized results.

V. CONCLUSION

This paper reaches the influence of multi-scale regions on pedestrian re-identification deeply, and proposes a posture-guided multi-scale structure relationship learning video pedestrian re-identification method. This method extracts the sample frame with the highest image quality and the most complete spatial information from the redundant video information based on the reference posture to reduce the influence of occlusion, posture change and redundant information on re-identification. The confidence of the joint point, the distance and the angle between the joint points are used as the decision-making index for posture discrimination. Aiming at the problem of re-identification of pedestrians in local regions, this paper divides multi-scale regions based on pedestrian joint points, and the relationship model further calculates the relationship scores between local areas to obtain relationship weights. The graph convolutional network based on relationship weights extracts the structural relationship features of multi-scale regions. Through experiments on three public data sets, according to the method proposed in this paper, the Rank-1 recognition accuracy rates of 85.5% (iLIDS-VID), 94.7% (PRID 2011) and 90.2% (MARS) have been obtained. It proves that the re-identification performance of multi-scale regional structure relationship features in complex scenes is more superior, and the generalization of the model is stronger. Future research will involve the evaluation of more challenging data sets, including related research on the issue of mutual occlusion of multiple pedestrians and the extraction of motion features in the case of pedestrian posture changes.

REFERENCES

[1] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato, "Hierarchical Gaussian descriptor for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1363–1372.

[2] J. Wang, Z. Wang, C. Liang, C. Gao, and N. Sang, "Equidistance constrained metric learning for person re-identification," *Pattern Recognit.*, vol. 74, pp. 38–51, Feb. 2018.

[3] Z. Wang, R. Hu, C. Chen, Y. Yu, J. Jiang, C. Liang, and S. Satoh, "Person reidentification via discrepancy matrix and matrix metric," *IEEE Trans. Cybern.*, vol. 48, no. 10, pp. 3006–3020, Oct. 2018.

[4] Y. Liu, J. Yan, and W. Ouyang, "Quality aware network for set to set recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5790–5799.

[5] Y. Wu, O. E. F. Bourahla, X. Li, F. Wu, Q. Tian, and X. Zhou, "Adaptive graph representation learning for video person re-identification," *IEEE Trans. Image Process.*, vol. 29, pp. 8821–8830, Jun. 2020.

[6] J. Yang, W.-S. Zheng, Q. Yang, Y.-C. Chen, and Q. Tian, "Spatial-temporal graph convolutional network for video-based person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3289–3299.

[7] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1302–1310.

[8] J. Yin, A. Wu, and W. S. Zheng, "Fine-grained person re-identification," *Int. J. Comput. Vis.*, vol. 128, no. 12, pp. 1654–1672, 2020.

[9] C. Patruno, R. Marani, G. Cicirelli, E. Stella, and T. D'Orazio, "People re-identification using skeleton standard posture and color descriptors from RGB-D data," *Pattern Recognit.*, vol. 89, pp. 77–90, May 2019.

[10] H. Cai, Z. Wang, and J. Cheng, "Multi-scale body-part mask guided attention for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1–10.

[11] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 274–282.

[12] Y. Fu, Y. Wei, Y. Zhou, H. Shi, G. Huang, X. Wang, Z. Yao, and T. Huang, "Horizontal pyramid matching for person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 8295–8302.

[13] X. Zhang, H. Luo, X. Fan, W. Xiang, Y. Sun, Q. Xiao, W. Jiang, C. Zhang, and J. Sun "AlignedReID: Surpassing human-level performance in person reidentification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Nov. 2017, pp. 1–10.

[14] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 480–496.

[15] H. Yao, S. Zhang, R. Hong, Y. Zhang, C. Xu, and Q. Tian, "Deep representation learning with part loss for person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2860–2871, Jun. 2019.

[16] F. Yang, K. Yan, S. Lu, H. Jia, X. Xie, and W. Gao, "Attention driven person re-identification," *Pattern Recognit.*, vol. 86, pp. 143–155, Feb. 2019.

[17] Y. Jing, C. Si, J. Wang, W. Wang, L. Wang, and T. Tan, "Pose-guided joint global and attentive local matching network for text-based person search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Oct. 2019, pp. 4321–4330.

[18] G. Song, B. Leng, Y. Liu, C. Hetang, and S. Cai, "Region-based quality estimation network for large-scale person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 1–8.

[19] Y. Liu, Z. Yuan, W. Zhou, and H. Li, "Spatial and temporal mutual promotion for video-based person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 8786–8793.

[20] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan, "End-to-end comparative attention networks for person re-identification," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3492–3506, Jul. 2017.

[21] L. Wu, Y. Wang, L. Shao, and M. Wang, "3-D PersonVLAD: Learning deep global representations for video-based person reidentification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 11, pp. 3347–3359, Nov. 2019.

[22] G. Chen, J. Lu, M. Yang, and J. Zhou, "Spatial-temporal attention-aware learning for video-based person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4192–4205, Sep. 2019.

[23] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2285–2294.

[24] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, and X. Wang, "HydraPlus-Net: Attentive deep features for pedestrian analysis," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 350–359.

[25] B. Chen, W. Deng, and J. Hu, "Mixed high-order attention network for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 371–381.

[26] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person re-identification by video ranking," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 688–703.

[27] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," in *Proc. Scand. Conf. Image Anal.*, 2011, pp. 91–102.

[28] 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 *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 868–884.

[29] W. Huang, C. Liang, Y. Yu, Z. Wang, W. Ruan, and R. Hu, "Video-based person re-identification via self paced weighting," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 2273–2280.
[30] M. Ye, J. Li, A. J. Ma, L. Zheng, and P. C. Yuen, "Dynamic graph co-matching for unsupervised video-based person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2976–2990, Jun. 2019.
[31] X. Zhu, X.-Y. Jing, X. You, X. Zhang, and T. Zhang, "Video-based person re-identification by simultaneously learning intra-video and inter-video distance metrics," *IEEE Trans. Image Process.*, vol. 27, no. 11, pp. 5683–5695, Nov. 2018.
[32] Z. Zhou, Y. Huang, W. Wang, L. Wang, and T. Tan, "See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6776–6785.
[33] S. Xu, Y. Cheng, K. Gu, Y. Yang, S. Chang, and P. Zhou, "Jointly attentive spatial-temporal pooling networks for video-based person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4733–4742.
[34] W. Zhang, X. He, W. Lu, H. Qiao, and Y. Li, "Feature aggregation with reinforcement learning for video-based person re-identification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 12, pp. 3847–3852, Dec. 2019.
[35] J. Li, S. Zhang, and T. Huang, "Multi-scale 3D convolution network for video based person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 8618–8625.
[36] C. Gao, Y. Chen, J.-G. Yu, and N. Sang, "Pose-guided spatiotemporal alignment for video-based person re-identification," *Inf. Sci.*, vol. 527, pp. 176–190, Jul. 2020.
[37] D. Chen, H. Li, T. Xiao, S. Yi, and X. Wang, "Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1169–1178.
[38] H. Liu, Z. Jie, K. Jayashree, M. Qi, J. Jiang, S. Yan, and J. Feng, "Video-based person re-identification with accumulative motion context," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 2788–2802, Oct. 2018.
[39] L. Bao, B. Ma, H. Chang, and X. Chen, "Preserving structural relationships for person re-identification," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2019, pp. 120–125.

DAN WEI received the B.S., M.S., and Ph.D. degrees from Hunan University, China, in 2006, 2008, and 2012. She is currently the Master Tutor with the Department of Mechanical and Automotive Engineering, Shanghai University of Engineering Science. Her research interests include person re-identification, intelligent transportation, and pattern recognition.

XIAOQIANG HU received the B.S. degree from the Anhui University of Agriculture and Economics, China, in 2018. He is currently pursuing the M.S. degree with the Department of Mechanical and Automotive Engineering, Shanghai University of Engineering Science. His research interests include person re-identification and artificial intelligence.

ZIYANG WANG received the B.S. degree from Xi'an Siyuan University, Xian, China, in 2016. He is currently pursuing the M.S. degree with the Department of Mechanical and Automotive Engineering, Shanghai University of Engineering Science. His research interests include person re-identification, artificial intelligence, and intelligent driving.

JIANGLIN SHEN received the B.S. degree from the Shanghai University of Engineering Science, China, in 2019. He is currently pursuing the M.S. degree with the Department of Mechanical and Automotive Engineering, Shanghai University of Engineering Science. His research interests include person re-identification and intelligent driving.

HONGJUAN REN received the B.S. and M.S. degrees from Shandong University, China, in 2000 and 2012, respectively, and the Ph.D. degree from Tongji University, Shanghai, in 2017. She is currently the Master Tutor with the Department of Mechanical and Automotive Engineering, Shanghai University of Engineering Science. Her research interests include automotive powertrain design and intelligent driving.