Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.061421

ARTICLE

Deep Learning Algorithm for Person Re-Identification Based on Dual Network Architecture

Meng Zhu[1,2], Xingyue Wang[3], Honge Ren[3,4,*], Abeer Hakeem[5] and Linda Mohaisen[5,*]

[1]College of Information Engineering, Harbin University, Harbin, 150086, China
[2]Heilongjiang Provincial Key Laboratory of the Intelligent Perception and Intelligent Software, Harbin University, Harbin, 150086, China
[3]College of Computer and Control Engineering, Northeast Forestry University, Harbin, 150040, China
[4]Heilongjiang Forestry Intelligent Equipment Engineering Research Center, Northeast Forestry University, Harbin, 150040, China
[5]Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, 21589, Saudi Arabia

*Corresponding Authors: Honge Ren. Email: rhe@nefu.edu.cn; Linda Mohaisen. Email: lmohaisen@kau.edu.sa

Received: 24 November 2024; Accepted: 13 January 2025; Published: 16 April 2025

ABSTRACT: Changing a person's posture and low resolution are the key challenges for person re-identification (ReID) in various deep learning applications. In this paper, we introduce an innovative architecture using a dual attention network that includes an attention module and a joint measurement module of spatial-temporal information. The proposed approach can be classified into two main tasks. Firstly, the spatial attention feature map is formed by aggregating features in the spatial dimension. Additionally, the same operation is carried out on the channel dimension to form channel attention feature maps. Therefore, the receptive field size is adjusted adaptively to mitigate the changing person posture issue. Secondly, we use a joint measurement method for the spatial-temporal information to fully harness the data, and it can also naturally integrate the information into the visual features of supervised ReID and hence overcome the low resolution problem. The experimental results indicate that our proposed algorithm markedly improves the accuracy in addressing changing human postures and low-resolution issues compared with contemporary leading techniques. The proposed method shows superior outcomes on widely recognized benchmarks, which are the Market-1501, MSMT17, and DukeMTMC-reID datasets. Furthermore, the proposed algorithm attains a Rank-1 accuracy of 97.4% and 94.9% mAP (mean Average Precision) on the Market-1501 dataset. Moreover, it achieves a 94.2% Rank-1 accuracy and 91.8% mAP on the DukeMTMC-reID dataset.

KEYWORDS: Person reidentification; ReID; computer vision; self-attention; spatial-temporal information

1 Introduction

The misalignment of the body part is a key factor that affects the person re-identification (ReID) results. This misalignment can be attributed to two main factors: Firstly, pedestrians adopt various postures while walking. Secondly, due to the limitation of ReID technology, the same body parts of the same person may appear in different proportions across images. ReID has become a prevalent area of focus within the realm of autonomous video surveillance for various security-based applications [1]. It solves the problem of cross-camera person identification and retrieval. ReID technology is usually used in intelligent monitoring systems, and the emergence of convolutional neural networks (CNNs) brings a breakthrough in progress for ReID [2]. Recently, ReID technology based on CNN has achieved good results in obtaining higher-level semantic

information by stacking convolutional layers. However, due to the influence of occlusion, pedestrian posture, light, and other factors, ReID technology is facing many more challenges [1].

Recent studies attempted to enhance the performance of ReID by using human structure information and concluded that discriminative local visual features can be captured by ReID with the help of this information. For example, the multi-scale contrast pooling feature method [3] employed ResNet50 (Deep Residual Network-50) for multi-scale feature extraction from the images of pedestrians. The layer of global average pooling is employed to acquire strong discriminative contrast pooling features across multiple networks. Furthermore, the multi-scale generative adversarial network method [4] used low-resolution images to reconstruct high-resolution pedestrian images as input. This method integrated the image details of different scales effectively and enhanced the performance of the occluded person images. Other existing methods relied on semantic segmentation algorithms [5] or semantic background information [6] to improve the performance of ReID. The above methods tried to make use of body parts or multi-scale features. However, these methods only adjust the receptive field size manually, which fails to provide a resolution for the variation in bodily proportions observed within different images of a given individual.

In addition to using human structure information, some existing methods also tried to enhance the precision of ReID depending on spatial-temporal information. The two methods, video-based and video-based ReID methods [7,8], used spatial and temporal information between frames to learn spatially varying and interdependent visual features. However, these methods focused only on the representation of visual features rather than the spatial-temporal constraints across different cameras.

The key contributions of this paper are summarised as follows:

- We introduce an innovative dual network architecture to solve person re-identification issues. It involves spatial/channel attention and a joint measurement module for spatial-temporal information to overcome the difficulty of adjusting the receptive field and partially using the spatial-temporal information.
- We solve the problems of low image resolution and changing pedestrian posture in the ReID problem. The spatial/channel attention module can achieve the adaptive adjustment of the receptive field size by aggregating features in spatial dimension and channel dimension to form a spatial/channel attention feature map. The spatial-temporal information joint measurement module models the spatial-temporal information, which can integrate spatial and temporal information into visual features of supervised ReID and improve the problem caused by low image resolution.
- Experiments are carried out on the Market-1501, MSMT17, and DukeMTMC-reID datasets, which show outstanding performance compared with the existing methods.

2 Related Works

Recent deep learning-based person ReID approaches are based on visual feature representation. Some models attempted to design local or global feature methods or multi-scale feature fusion methods, such as the Singular Vector Decomposition Network (SVDNet) [9], MultiScale [10], Diversified Local Attention Network (DLAN) [11], Contrast [3], and multi-granularity feature fusion [12]. Some models used an attitude-based approach such as Latent Parts [13], Part-Aligned [14], and Generative Adversarial Network (GAN) hybrid coding [15]. Furthermore, other models tried to use spatial-temporal information such as TFusion-sup [16].

Huang et al. [17] introduced a lightweight network for person re-identification based on Multi-Scale Focus Attention called MSFANet, which leverages the attention mechanisms to acquire multi-scale feature representations. They introduced a function for the fusion loss that combines softmax and weighted TriHard

loss functions. Their model achieved better accuracy and computational complexity. Song et al. [18] introduced a new label assignment approach called Dynamic Adaptive Label Allocation (DALA), which relies on metric relationships among features only and doesn't necessitate clustering. They generated appropriate pseudo-labels by avoiding quantitative loss and offered enhanced flexibility. Additionally, they introduced an auxiliary attention module named CATA to enhance feature robustness, which was integrated into a convolutional neural network (CNN). Zhu et al. [19] proposed a framework for extracting more diverse feature representations by utilizing global differences between multi-granularity features. Batool et al. [20] presented a hybrid model called a pseudo-labeled Omni-Scale Network (POSNet) that tackles two important person ReID issues. They solved the issue of limited labeled data by giving pseudo-labels to unlabeled data. Furthermore, they tackled the issue of intra- and inter-class variations by employing an improved omni-scale feature learning approach that builds the labeled and unlabeled data feature space using temporal features and soft-pool attention.

Moreover, it is possible to record long-range relationships without sacrificing useful short-term temporal information. In order to gradually integrate frame-level characteristics, a novel Spatio-Temporal Aware Network (STAN) was proposed by Wang et al. [21]. A spatio-temporal complementary module was designed. The temporal interaction and spatial reference modules were two main parts of this module. Through the correlation of the feature and reference nodes, the module of temporal interaction was intended to suppress the features that aren't important and improve the discriminative quality of temporal features. To generate detailed information within the video frames at a granular level, they designed the spatial reference module. Additionally, by producing reference characteristics, a temporal attention module was intended to enhance the temporal interaction network. Behera et al. [22] proposed a Person Graph Attention-based Network (PGAN) using graph convolution and attention networks. They considered personal attributes and body parts as distinct points for building a graph.

Overall, the abovementioned methods are suffering from some challenges. These methods may take too long time to train because it has too many parameters. Furthermore, in certain situations, these methods might not be able to identify the target pedestrian correctly due to the subpar image quality resulting from the outdated camera and the mistake of the meticulous annotation. Additionally, these methods still cannot solve the problem of appearance ambiguity and cannot make full use of spatial-temporal information. Compared with the above-mentioned methods, the proposed algorithm can locate the body parts in the input image under different poses and proportions. Additionally, the consolidation of these features enhances the neural network's capability for feature extraction. Simultaneously, the spatial-temporal information is modeled to filter irrelevant images and it can additionally improve the model's performance.

3 Proposed Approach and Model Architecture

3.1 Attention Mechanism

This paper presents a Spatial-Channel Network (SCNet) based on Spatial Attention (SA) and Channel Attention (CA) to improve the feature representation ability of CNN networks. Different from the traditional CNN, which extracts the features of fixed geometric structure, our proposed network can adjust the receptive field according to the posture of the person and the body parts proportion of the input images. More specifically, by generating spatial semantic relationship mapping, it is found that there are two types of interdependence between different locations of the image: appearance relationship with higher correlation for locations that have similar features and location relationships where locations are in close proximity to each other and demonstrate heightened correlations. In such a manner, body parts with multiple positions and proportions can adjust the receptive field size automatically. Based on the spatial relationship map, the feature map is updated by aggregating the semantic information at various positions.

Additionally, we employ a self-attention mechanism in the channel to model the semantic correlation between the channels and enhance the feature extraction capacity of CNN, especially for small-scale visual information that is easy to disappears in advanced features extracted from CNN (such as luggage, etc.). Furthermore, the channel attention can aggregate the semantic similarity of the vision clues in all channels.

3.1.1 Spatial Self-Attention Mechanism

Firstly, F is reshaped to $\mathbb{R}^{C\times M}$, as indicated in Fig. 1, where M $(M = H \times W)$ is the number of spatial features, and then the interconnectivity among spatial features is simulated to produce a graph of semantic relationships. Considering the relationship between appearance information and location information, we aggregate the associated spatial features in the aggregation operation by utilizing the generated semantic relationship diagram.

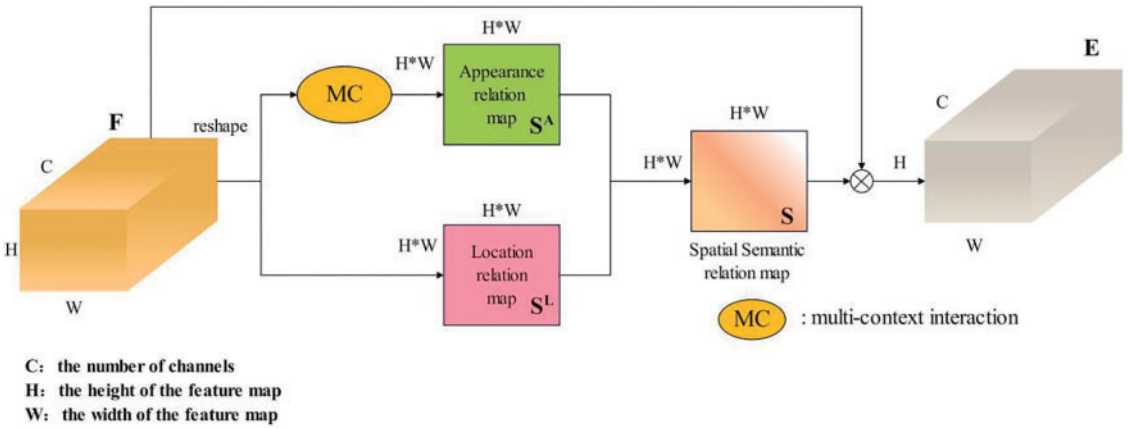

Figure 1: The structure of the spatial attention (SA) module

The local features of adjacent locations possess a high degree of correlation, as their receptive fields frequently overlap. Translation: Due to overlapping receptive fields at adjacent locations, there is a higher local similarity between them; windows at adjacent locations can capture more accurate appearance features. In this paper, contextual information is used to obtain more accurate appearance-similar features.

As shown in Fig. 2, we extract the $K \times K$ regions surrounding points i and j within P_i and P_j, respectively, to determine the similarity in appearance between F1 and F2, and the dot product results between the corresponding location features are accumulated. All the spatial locations of F are normalized to obtain the appearance similarity in softmax as follows:

$$\left(S_K^A\right)_{ij} = \frac{\exp\left(\sum_{k=1}^{K\times K}\left(p_{i,k}^T p_{j,k}\right)\right)}{\sum_{t=1}^{H\times W}\exp\left(\sum_{k=1}^{K\times K}\left(p_{i,k}^T p_{t,k}\right)\right)} \tag{1}$$

where $p_{i,k}$ and $p_{j,k}$ indicate the features of P_i and P_j at the kth spatial position. Generally, softmax suppresses the features corresponding to the small similarity of different body parts significantly. By fusing contextual information and suppressing differences, the diagram can be used to locate body parts in different poses and scales roughly. Since only one window size (i.e., $k = 1$) is considered, S^A is a single-context appearance relationship mapping. The multi-context appearance relational mapping S_K^A is calculated as in Eq. (2):

$$S^A = softmax\left(\mathcal{F}\left(S_1^A, \ldots S_N^A\right)\right), \tag{2}$$

where F signifies the fusion function of the element product and N represents the number of context levels.

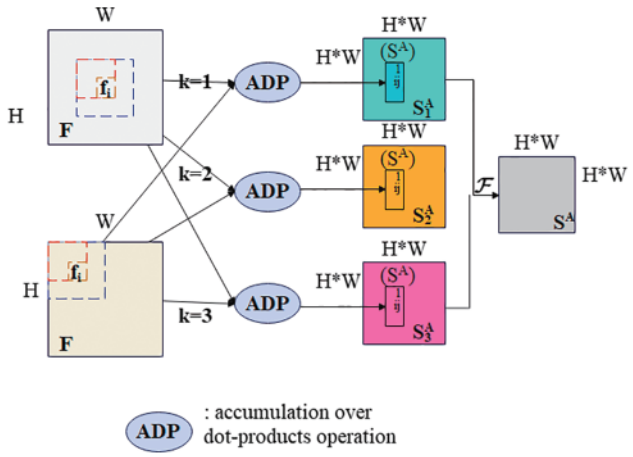

Figure 2: Multi-context interaction operation of spatial attention (SA)

The local features of the same body parts are similar in space. Consequently, this paper presents the integration of locational relationships, leveraging spatial structural information more comprehensively alongside the features that are near a certain area that exhibit an augmented correlation. In a formal manner, the position relationship between spatial features f_i and f_j is calculated using the following formula:

$$l_{ij} = \frac{1}{2\pi\sigma_1\sigma_2} \exp\left[-\frac{1}{2}\left(\frac{\left(x_j - x_i\right)^2}{\sigma_1^2} + \frac{\left(y_j - y_i\right)^2}{\sigma_2^2}\right)\right] \tag{3}$$

where (x, y) indicate the position coordinates for the feature, (σ_1, σ_2) are the standard deviations that can be used to adjust the Gaussian function. Normalize l_{ij}, the aggregate of the associative position f_i relation values are normalized to equal 1. The spatial location relation diagram S^L can be expressed as follows:

$$\left(S^L\right)_{ij} = \frac{l_{ij}}{\sum_{t=1}^{H\times W} l_{it}} \tag{4}$$

The positional relationship between f_i and f_j decay exponentially as their spatial separation increases. Moreover, the algorithm S^L can help capture more accurate visual information.

Spatial semantic relation combines appearance similarity with location relation, which can be expressed as follows:

$$S = softmax\left(\mathcal{F}\left(S^A, S^L\right)\right) \tag{5}$$

The spatial features obtained based on the semantic relationship graph are then aggregated. The aggregated feature map $E^s \in \mathbb{R}^{C\times M}$ is computed as follows:

$$E^S = FS^T \tag{6}$$

3.1.2 Channel Attention

Current person ReID models usually superimpose multiple convolutional layers to get semantic information of higher levels. When the layer count escalates, these models run the risk of forfeiting intricate

visual details. However, fine-grained features are important to distinguish between two pedestrians with small differences between classifications.

Most of the existing channel maps with advanced feature fusion showed strong responses to specific parts. Therefore, the present paper introduces a channel attention module aimed at consolidating semantically analogous features within all channels. This module serves to elevate the feature representation capabilities of neural networks when dealing with specific components. The main structure of the Channel Attention module (CA module) is presented in Fig. 3. From the interaction phase, by meticulously capturing the inherent semantic correlations among diverse channels through the employment of CA, a remarkable channel semantic relationship map is produced for the feature map F. To do this, the first step is to reshape F to $\mathbb{R}^{C\times M}$, where M is the product of H and W. Then, matrix multiplication is applied to the transpose of F and F. Then, the result is meticulously normalized, yielding the exquisite channel semantic relation map $C \in \mathbb{R}^{C\times C}$. By applying the following formula, we can calculate the semantic similarity between any two channels:

$$C_{mn} = \frac{\exp\left(f_m^T f_n\right)}{\sum_{t=1}^{C} \exp\left(f_m^T f_l\right)} \tag{7}$$

where $f_m, f_l \in \mathbb{R}^M$ are the features in mth and nth channels of the F.

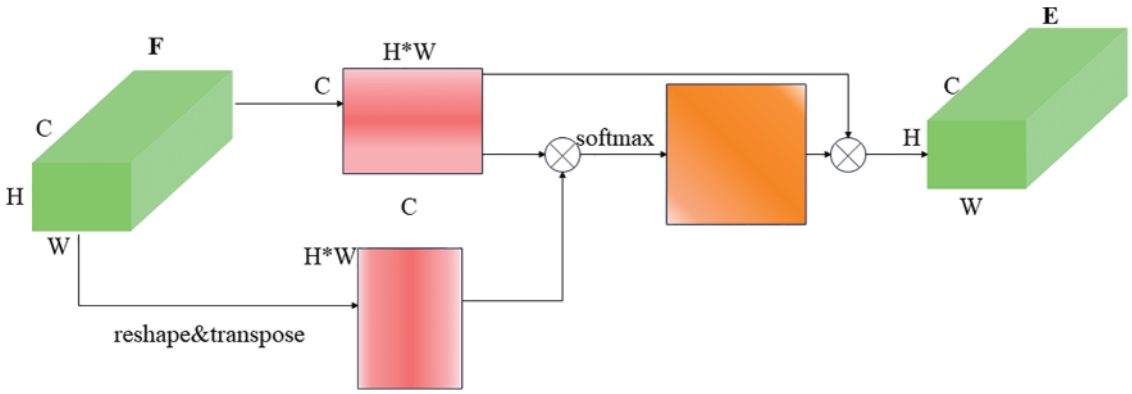

Figure 3: The main architecture of channel attention (CA) module

In the following aggregation operation, we compile the channel features by the mapping of channel relationships, and matrix multiplication is applied on C and F to obtain aggregated feature map $E^C \in \mathbb{R}^{C\times M}$ where

$$E^C = CF \tag{8}$$

Hence, to keep the same input size, we reshape E^C to $\mathbb{R}^{C\times H\times W}$. It is noteworthy that, the generated feature map effectively aggregates features that are semantically similar by using an input-specific channel relational graph C. Therefore, SA aggregates features based on spatial relationship graphs, while CA aggregates features based on channels. Similarly, CA also can adjust the input feature map adaptively as in SA, which proves advantageous in enhancing the capability of convolutional neural networks for feature extraction.

3.1.3 Model Aggregation

Now, we combine the attention module into the existing architecture. As shown in Fig. 4a, the SC block can be addressed as follows:

$$Y = BN(E) + F \tag{9}$$

where E is the output of the SA or CA module as given in Eq. (5) or Eq. (7). F is the inputted feature maps. The $(+F)$ represents residual connection, which can ensure the network's performance to a maximum extent.

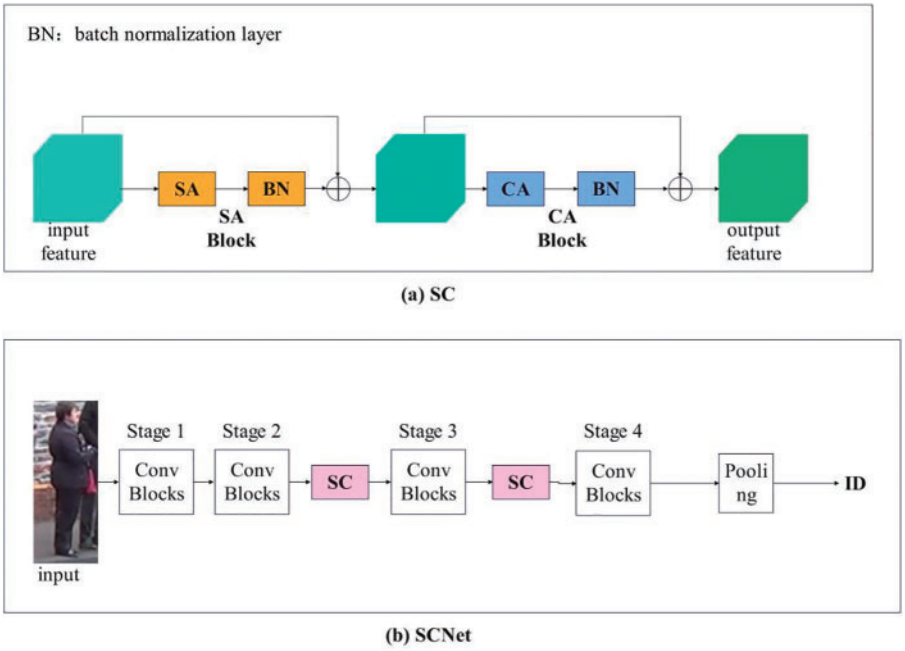

Figure 4: (a) The architecture of the SC block, consisting of SA and CA blocks in turn; (b) Inserting SC module into the network

3.2 Spatial-Temporal Information

The majority of the existing person ReID methods ignore the spatial-temporal information. Consequently, if the amount of data is huge, the traditional methods will be disturbed by the ambiguity problem of pedestrian representation under cross-camera and cannot achieve good performances. In this paper, a spatial-temporal information-based module (ST) is proposed to explore the spatial-temporal information of images. To model an intricate probability distribution of the spatial-temporal, we employ Histogram-Parzen (HP) approach to obtain spatio-temporal information. Furthermore, we assist the visual feature stream by capturing the spatial-temporal auxiliary information.

In this work, we estimate the spatial-temporal histogram using the Histogram-Parzen method and then smooth the histogram using the Parzen window method. This helps to alleviate the computational burden of the non-parametric estimation method (the Parzen window method). The (ID_i, C_i, t_i) and (ID_j, C_j, t_j), $(t_i < t_j)$ denote the identity label, camera ID, and time stamp for the two images I_i and I_j, respectively. Then, we depict the likelihood of a positive-image pair by creating a rough spatial-temporal histogram using the

following formula:

$$\hat{p}\left(y=1|k,ci,cj\right)=\frac{n_{cicj}^{k}}{\sum_{l}n_{cicj}^{l}} \tag{10}$$

where k represents the kth column of the histogram. The parameter $n_{C_iC_j}^{k}$ denotes the number of image pairs depicting individuals whose time difference from C_i to C_j is on the kth column. The time difference $tj-ti\in\left(\left(k-1\right)\Delta t,k\Delta t\right)$. Hence, $y=1$ means that I_i and I_j share the same human body ID_i (i.e., $ID_i=ID_j$), while y = 0 represents different identities ID ($ID_i\neq ID_j$).

The histogram is smoothed by using the following formula:

$$\hat{p}\left(y=1|k,ci,cj\right)=\frac{1}{Z}\sum_{l}\hat{p}\left(y=1|l,ci,cj\right)K\left(l-k\right) \tag{11}$$

where $K\left(.\right)$ is the kernel. The factor $Z=\sum_{k}p\left(y=1|l,ci,cj\right)$ represents the normalization factor. The Gaussian function is used as the kernel as follows:

$$K\left(x\right)=\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-x^2}{2\sigma^2}} \tag{12}$$

ST module can exclude many irrelevant images in the gallery to narrow the scope of the query.

3.3 Network Structure

The objective of the spatial-channel (SC) approach is to harness the similarity of the visual features and spatio-temporal constraints within a consolidated framework. Therefore, in this paper, we propose to use a dual-flow structure composed of three submodules, including a visual feature flow, a spatial-temporal flow, and a joint measurement module, as shown in Fig. 5. Fig. 5 shows the dual-stream architecture of the proposed SCNet.

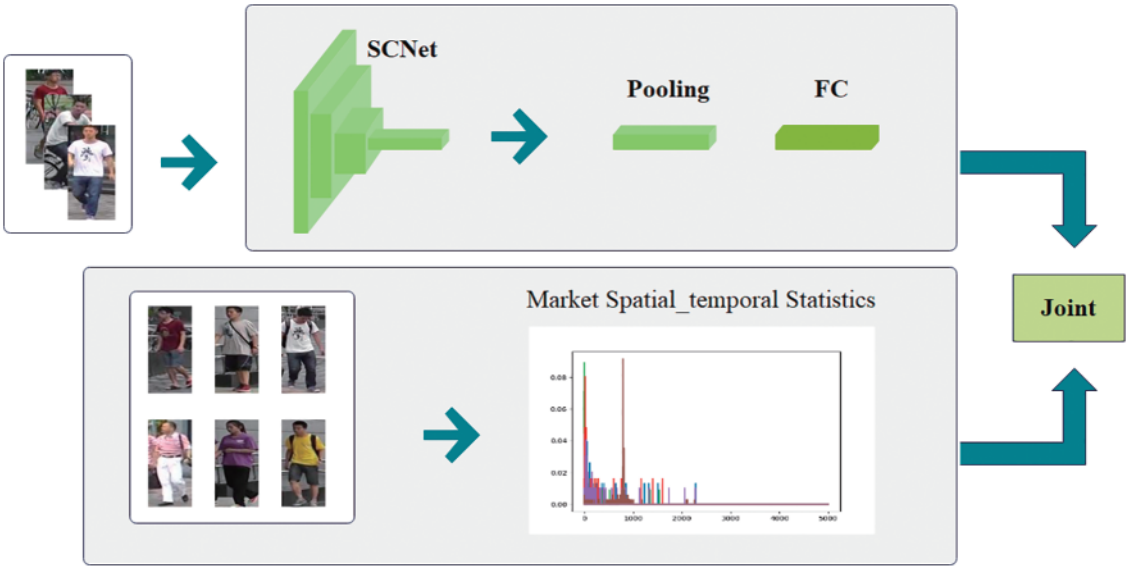

Figure 5: Dual stream architecture for the proposed SCNet

A collective similarity evaluation metric incorporating Logistic Smoothing (LS) is used for incorporating disparate forms of information into a consolidated network through a composite similarity metric. Hence, to obtain two distinct patterns. Joint Probability is expressed in Eq. (13) as follows:

$$p\left(y=1|xi,xj,k,ci,cj\right)=s\left(xi,xj\right)p\left(y=1|k,ci,cj\right) \tag{13}$$

There are two issues with Eq. (13). Firstly, it is deemed impractical to directly equate the similarity score with the probability of visual occurrence. Secondly, the probability of the spatio-temporal $p\left(y=1|k,ci,cj\right)$ is neither reliable nor controllable; this is due to the uncertainty surrounding individuals' walking patterns and speed. The spatial-temporal probability function $p\left(y=1|k,ci,cj\right)$ used for maintaining the same accuracy will lead to lower recall. For an image that needs to be queried, if one target image has a high appearance similarity but a spatial-temporal probability of only 0.01, while another image has a similarity score of 0.4 and a spatial-temporal probability of 0.1. Eq. (13) tends to return the second image.

Furthermore, there exists the issue of low spatial-temporal probabilities caused by pedestrians moving too fast. To address this challenge, we recommend the implementation of a logistic smoothing method. In literature, the Logistic Models (LM) are widely used for binary classification problems. It is expressed as in Eq. (14) as follows:

$$f\left(x;\lambda,\gamma\right)=\frac{1}{1+\lambda e^{-\gamma x}} \tag{14}$$

where λ and γ are constants, λ represents a smoothing factor, and γ indicates the contraction factor.

The proposed logistic smoothing method in this paper not only adjusts the probability of low-probability events but also calculates the probability that two images with specific information belong to the same identity. Hence, we modify Eq. (13) to be as follows:

$$p_{joint}=f\left(s;\lambda_0,\gamma_0\right)f\left(p_{st};\lambda 1,\gamma 1\right) \tag{15}$$

To simplify notation, p_{joint}, s and p_{st} are used in this paper to represent $p\left(y=1|xi,xj,k,ci,cj\right)$, $s\left(xi,xj\right)$ and $p\left(y=1|k,ci,cj\right)$; respectively. From Eq. (9), it can be noted that $s\in(-1,1)$ is compressed by logic functions at that time, just like Laplacian smoothing, but the degree is not so great. In contrast to this, $pst\in(0,1)$ is truncated and boosted substantially. Even the space-time pst is close to zero probability $f\left(pst;\lambda 1,\gamma 1\right)\geq f\left(0\right)=\frac{1}{1+\lambda 1}$. Since, as mentioned above, spatial-temporal probabilities are unreliable and visual similarity is relatively reliable. So, Eq. (15) is robust to low-probability events by logical smoothing.

4 Experimental Results and Analysis

4.1 Dataset Settings and Metrics

In this study, the effectiveness of the proposed approach is assessed through the evaluation of three extensive datasets for pedestrian re-identification: Market-1501 [23], DukeMTMC-reID [24], and the MSMT17 dataset. We perform ablation experiments to measure the performance improvement for each module and compare our algorithm with some existing state-of-the-art (SOTA) methods. The evaluation metrics used in this research are mean Average Accuracy (mAP) and Rank-1. While mAP measures the overall performance, Rank-1 represents the probability of correctly identifying the first target.

The Market-1501 [23] dataset was established near a supermarket located at Tsinghua University, utilizing a combination of five high-resolution digital cameras and one low-resolution device. It comprises a vast collection of 32,668 annotated images featuring 1501 pedestrians. Multiple cameras were employed to

capture images of each individual in the dataset, ensuring comprehensive coverage. Additionally, each image includes relevant information such as frame number and camera ID details.

DukeMTMC-reID [24] dataset is a subset of the DukeMTMC dataset. It has 1404 pedestrians appearing in more than 2 cameras and 408 pedestrians appearing in only 1 camera.

4.2 Experimental Environment and Parameters Configuration

The experimental processes were conducted using a consistent software and hardware environment to ensure result consistency. The experiments were conducted on an Ubuntu 21.10 operating system, with hardware specifications including 32 GB of memory, an Intel I7-10700K CPU, an NVIDIA RTX3090 graphics card, and software versions such as CUDA11 for GPU acceleration, Python3.6 for programming, and PyTorch1.6 for deep learning framework.

This paper employs the PCB (Part-based Convolutional Baseline) [25] after adding the SC module network to generate a stream of visual features. During the training, the image size is uniformly adjusted to 288 × 144, and it is randomly cropped to 256 × 128. During the training process, the batch size is set to 32, and the Adam optimizer is selected to optimize the proposed model, and 60 epochs in total are applied. The learning rate is adjusted to 0.1, which is then minimized by a factor of 10 to 0.01 following the completion of 40 training epochs. We employ a cross-entropy loss function. After the SC module is added to the 2nd and 3rd residual blocks of ResNet, the number of context levels is set to 3. In the SC, after the second and third residual blocks, the σ_1 is set to 10 and 5, respectively, while σ_2 is set to 20 and 10, respectively. For spatio-temporal information flow, the time interval Δt is set to a total of 100 frames, and the Gaussian kernel parameter σ is set to a value of 50. For the joint measurement, we set $\lambda 1$, $\lambda 2$, $\gamma 1$ and $\gamma 2$ to 1, 1, 5, and 5, respectively.

For the joint metric module, we conduct some experiments to evaluate the impact of the parameter smoothing factor λ and contraction factor γ on the model performance. The experimental results, as shown in Fig. 6, indicate that when the smoothing factor λ is within the range of 0.8 to 1.4 and the contraction factor γ is within the range of 3 to 5, the spatio-temporal information flow can ensure the best performance. Experiments demonstrate that the smoothing factor λ and contraction factor γ are robust to the model performance, and the model can still maintain a high level of performance even within a specific range of variations.

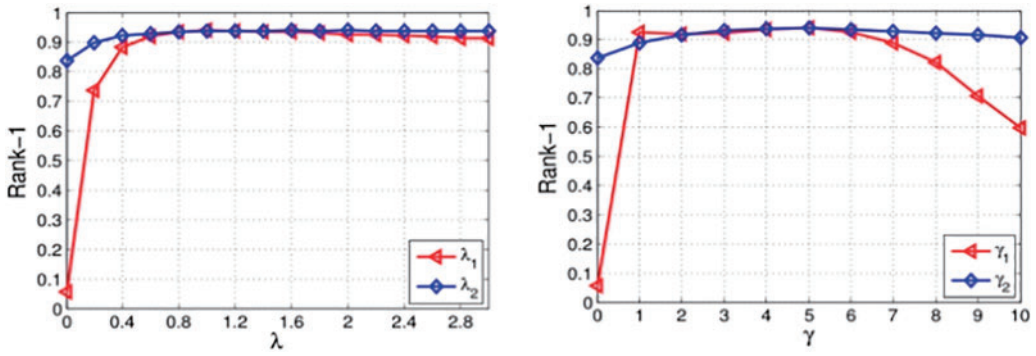

Figure 6: The impact of the hyperparameters on the model's performance

4.3 Ablation Studies

To verify the impact of each module in the proposed approach, multiple ablation studies are performed on the DukeMTMC-reID, MSMT17, and Market-1501 datasets, as listed in Table 1. As we can see, in the second row, visual features are extracted by PCB combined with spatial channel SC dual attention (PCB + SC). In the third row, the spatial-temporal information joint measurement method is used to identify pedestrians by (PCB + ST). The fourth row proposes two modules of dual attention and joint measurement of spatial-temporal information in the spatial-temporal channel (PCB + SC + ST).

Table 1: Comparison results of the ablation experiment

Algorithm	DukeMTMC-reID dataset		Market-1501 dataset	
	Rank-1 (%)	**mAP (%)**	**Rank-1 (%)**	**mAP (%)**
PCB [25]	81.9	65.3	92.4	77.3
PCB + SC	88.0	74.1	95.7	84.3
PCB + ST	92.14	80.2	96.4	85.8
PCB + SC + ST (Ours)	**94.2**	**91.8**	**97.2**	**86.5**

As shown in Table 1, integrating the SC module into the conventional CNN architecture of the PCB network yielded significant improvements in Rank-1 and mAP metrics. In the DukeMTMC-reID dataset, the Rank-1 accuracy increased by 6.1%, and mAP increased by 8.8%. Similarly, on the Market-1501 dataset, the Rank-1 accuracy experienced a boost of 3.3%, while the mAP showed an improvement of 7.0%. Additionally, from DukeMTMC-reID, Rank-1 is enhanced by 10.24% and the mAP is enhanced by 14.9%, whereas on the Market-1501 dataset, Rank-1 is increased by 4.0% and the mAP is enhanced by 8.5% when the spatio-temporal module (ST) is inserted into the traditional CNN of the PCB network. Finally, the combined integration of the ST and SC modules into the PCB network yielded substantial improvements in both Rank-1 and mAP metrics. On the DukeMTMC-reID dataset, the Rank-1 accuracy showed a remarkable increase of 12.3%, accompanied by an impressive mAP improvement of 26.5%. Similarly, on the Market-1501 dataset, the Rank-1 experienced a notable boost of 4.8%, while the mAP exhibited a significant improvement of 9.2%.

To evaluate the impact of the SC (Spatial Channel) module on feature extraction, a comparative analysis is conducted, and CAM (Channel Attention Mechanism) is used to visualize the image feature receptive fields learned by SC, as illustrated in Figs. 7 and 8. The SC module is defined as the effective receptive fields with higher correlation. It can be observed that SC can adjust the size and range of the receptive field in different situations adaptively. To verify the effect of the SA and CA combination on the accuracy of the SC module, in this paper, we compare three different arrangement methods: parallel, serial channel space, and serial space channel. The results are shown in Table 2, which indicates that the combination of SA + CA has a higher accuracy.

Furthermore, to precisely evaluate the role of the Spatial Semantic Attention (SSA) module and Channel Semantic Attention (CSA) module in enhancing the person re-identification accuracy, we conduct detailed and in-depth ablation studies on the DukeMTMC-reID and Market-1501 datasets. Using the ResNet50 network as a reference baseline, the SSA module was first integrated into the baseline model to measure its contribution in improving the performance of the model. Subsequently, to evaluate the effectiveness of the channel attention mechanism on the performance, the SSA module is replaced with the CSA module. After evaluating each module individually, we further explore the comprehensive impact of integrating both the SSA and CSA modules into the baseline model, under their synergistic effect. In this way, the aim is to deeply

understand the independent and combined effects of these attention modules on enhancing the accuracy of person re-identification tasks.

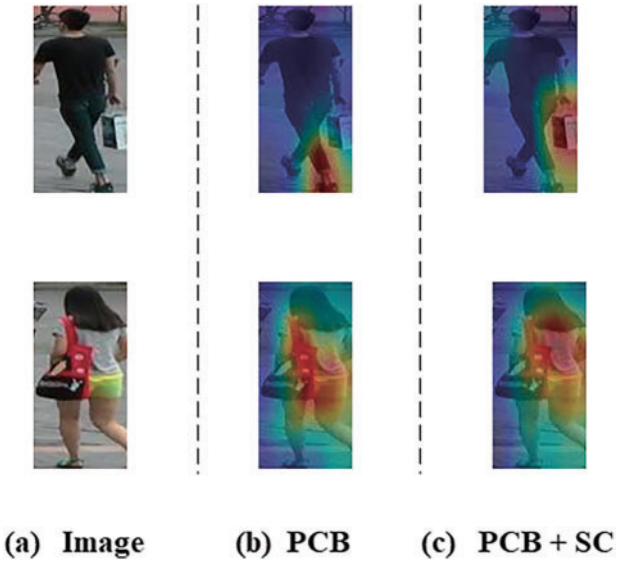

Figure 7: Some visualization results of the proposed SC module from the Market-1501 dataset

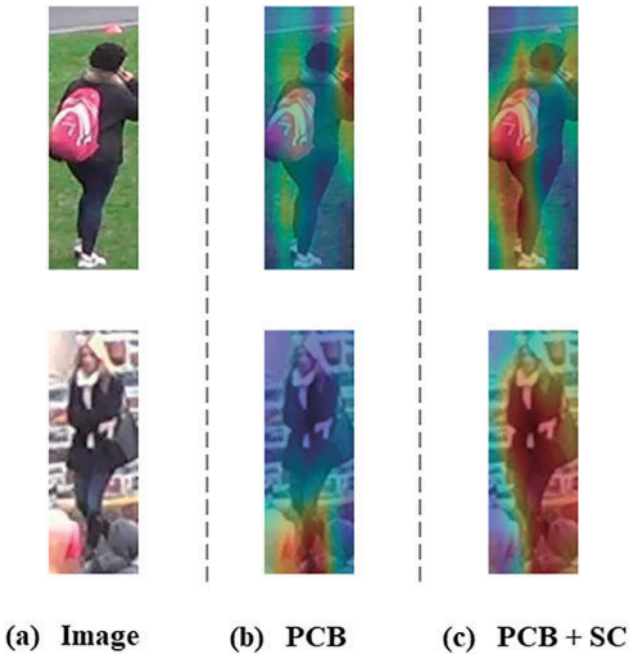

Figure 8: Visualization results of the proposed SC module from DukeMTMC-reID dataset

Table 2: Results of SA and CA combination methods on different datasets

Combination method	DukeMTMC-reID dataset		Market-1501 dataset		MSMT17 dataset	
	Rank-1 (%)	mAP (%)	Rank-1 (%)	mAP (%)	Rank-1 (%)	mAP (%)
PCB [25]	81.9	65.3	92.4	77.3	75.8	52.3
PCB + SA & CA	86.4	73.1	94.3	83.6	77.2	60.5
PCB + CA + SA	86.7	73.1	94.6	83.6	78.5	63.0
PCB + SA + CA (Ours)	**86.9**	**73.1**	**94.8**	**84.3**	**80.2**	**64.4**

Table 3 shows the specific impact of the SSA and CSA modules on the accuracy. The data indicates that performance significantly improves when these two modules are used in conjunction. On the DukeMTMC-reID dataset, when both modules are active, Rank-1 reaches 86.9%, and the mAP reaches 74.1%. On the Market-1501 dataset, Rank-1 accuracy increased to 94.8%, and mAP improved to 84.3%. These results are superior to the performance when only the SSA or CSA module is applied individually to the baseline model.

Table 3: The ablation study comparison results of the SSA and CSA on the DukeMTMC-reID, and Market-1501 dataset

SSA	CSA	Market-1501 dataset		DukeMTMC-reID dataset	
		Rank-1 (%)	mAP (%)	Rank-1 (%)	mAP (%)
		90.4	76.2	82.1	66.0
√		93.7	84.3	81.0	71,1
	√	91.4	82.5	83.4	70.2
√	√	94.8	84.3	86.9	74.1

4.4 Comparison Experiments

The ablation studies and comparison experiments of the proposed model are performed on the Market-1501 and DukeMTMC-reID datasets. Comparisons with some existing SOTA techniques are conducted to show the advantages and efficiency of our model. The proposed person ReID model is evaluated against 15 existing approaches, which can be classified into six categories: 1) Manual feature-based approaches: including BoW + Kissmeh [26] and the Null Space [27]; 2) Global feature-based methods: such as SVDNet [9] and improvement based on global features pedestrian re-recognition method [28]; 3) Local feature-based methods: including MultiScale [10], PSE + ECN [29], PDC [30] and DLAN [11]; 4) Models based on multi-scale fusion: such as Contrast [3] and multi-granularity feature fusion methods [12]; 5) Posture-based methods: including Latent Parts [13], Part-Aligned [14], PCB [25] and GAN network hybrid coding methods [15]; 6) Spatial-temporal methods: including TFusion-sup Method [16].

The data depicted in Table 4 illustrate that our proposed approach outperforms the existing prominent pedestrian ReID techniques. Specifically, on the Market-1501 dataset, we achieve a Rank-1 of 97.4% and a mAP of 94.9%; on the DukeMTMC-reID dataset, Rank-1 accuracy reaches 94.2%, with a mAP of 91.8%. The SVDNet, PSE + ECN, MultiScal, TFusion-sup, and other outstanding methods have been improved to different degrees. To further verify the effectiveness and accuracy of the proposed method, we conducted some experiments on the MSMT17 dataset. Compared with the traditional datasets such as Market-1501 and DukeMTMC-reID, the dataset MSMT17 has a larger scale and presents more challenges, including a greater

number of identities, camera views, and complex real-world scenarios. On the MSMT17 dataset, the Rank-1 accuracy reaches 84.1%, with the mAP at 69.4%.

Table 4: Comparisons of the proposed method against existing methods on the Market-1501, DukeMTMC-reID, and MSMT17 datasets

Algorithm	Market-1501 dataset		DukeMTMC-reID dataset		MSMT17 dataset	
	Rank-1 (%)	mAP (%)	Rank-1 (%)	mAP (%)	Rank-1 (%)	mAP (%)
BoW + kissme [26]	44.4	20.8	25.1	12.2	–	–
Null Space [27]	55.4	29.9	–	–	–	–
SVDNet [9]	82.3	62.1	76.7	56.8	77.0	53.3
Literature [28]	94.4	92.6	90.9	88.6	83.8	65.8
MultiScale [10]	88.9	73.1	79.2	60.6	–	–
PDC [30]	84.1	63.4	–	–	–	–
PSE + ECN [29]	90.3	84.0	85.2	79.8	82.4	63.9
DLAN [11]	95.1	88.4	73.4	71.8	–	–
Contrast [3]	95.0	87.6	88.6	77.7	80.6	62.8
Literature [12]	93.7	82.4	85.8	75.1	–	–
Latent Parts [13]	80.3	57.5	–	–	–	–
Part-Aligned [14]	81.0	63.4	–	–	–	–
PCB [25]	91.2	75.8	83.8	69.4	78.7	52.9
Literature [15]	93.4	82.2	84.3	70.5	82.4	68.9
TFusion-sup [16]	73.1	–	–	–	–	–
Ours	**97.4**	**94.9**	**94.2**	**91.8**	**84.1**	**69.4**

In general, using our proposed method, the model can locate the body parts in the input image under different poses and proportions. The consolidation of these features enhances the neural network's capability for feature extraction. Additionally, the employed spatial-temporal information joint measurement method can use the information, such as camera ID and timestamp, to filter out the irrelevant pedestrian images effectively, thereby narrowing the query scope. Our proposed method can enhance performance effectively and has a certain generalisation ability compared to the existing methods. However, the accuracy of the spatial-temporal information still requires some improvements. In order to overcome these challenges, many techniques utilizing neural networks can be used to learn robust representations from various sources of information. In the future, more recent approaches such as visual-modified attention, multi-scale attention with transformer-CNN architecture, Graph Relearn Network, and other techniques can be considered to enhance the network performance.

5 Conclusion

Person re-identification is a significant and challenging task for many public security and video surveillance applications. The main issues with the person re-identification are low resolution and changing pedestrian posture. In this paper, we introduce a Spatial-Channel Network (SCNet) based on spatial attention and channel attention for person re-identification task. To fully utilize spatiotemporal details and adaptively adjust the receptive field size, we combine SCNet with the spatial-temporal information joint measurement approach and the channel spatial dual attention model. The proposed algorithm is evaluated on Market-1501, DukeMTMC-reID, and MSMT17 datasets to validate its performance. Furthermore, some ablation

experiments are conducted to verify the impact of each module in the proposed network in improving the overall performance. The proposed algorithm achieves a Rank-1 accuracy of 97.4% and mAP of 94.9% on the Market-1501 dataset. On the DukeMTMC-reID dataset, the proposed algorithm achieves a Rank-1 accuracy of 94.2% and mAP of 91.8%. Although the proposed algorithm is capable of successfully overcoming occlusion issues, pedestrian attitude changes, and low-resolution problems, some improvements are still needed. In future work, we will try to enhance the global information and modeling accuracy of the spatial-temporal information based on the existing methods. Furthermore, visual-modified attention, multi-scale channel-spatial attention mechanisms, transformer-based CNN architecture, and other techniques can be considered to improve the network performance.

Acknowledgement: The authors would like to thank the editorial office and reviewers for their valuable comments and suggestions to enhance the paper quality.

Funding Statement: This work was supported by the Young Doctoral Research Initiation Fund Project of Harbin University "Research on Wood Recognition Methods Based on Deep Learning Fusion Model" (Project no. HUDF2022110), the Self-Funded Project of Harbin Science and Technology Plan "Research on Computer Vision Recognition Technology of Wood Species Based on Transfer Learning Fusion Model" (Project no. ZC2022ZJ010027), and the Fundamental Research Funds for the Central Universities (2572017PZ10).

Author Contributions: Meng Zhu was responsible for model design, model training, code writing, and debugging. Xingyue Wang was responsible for model design, dataset construction, and code debugging. Honge Ren contributed to conceptualization, methodology, draft writing, reviewing, and provide experimental conditions, including the artificial intelligence laboratory and experiment equipment. Abeer Hakeem and Linda Mohaisen contributed to dataset annotation checking, article editing, supervision, investigation, and revision. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Ye M, Shen J, Lin G, Xiang T, Shao L, Hoi SCH. Deep learning for person re-identification: a survey and outlook. IEEE Trans Pattern Anal Mach Intell. 2022;44(6):2872–93. doi:10.1109/TPAMI.2021.3054775.
2. Ming Z, Zhu M, Wang X, Zhu J, Cheng J, Gao C, et al. Deep learning-based person re-identification methods: a survey and outlook of recent works. Image Vis Comput. 2022;119:104394. doi:10.1016/j.imavis.2022.104394.
3. Liu XR, Li XX, Qin CH. Person re-identification method with multi-scale contrast pooling feature. Comput Eng. 2022;48(4):292–8 (In Chinese). doi:10.19678/j.issn.1000-3428.0061508.
4. Yang WX, Yan Y, Chen S, Zhang XK, Wang HZ. Multi-scale generative adversarial network for person re-identification under occlusion. J Softw. 2020;31(7):1943–58 (In Chinese). doi:10.13328/j.cnki.jos.005932.
5. Kalayeh MM, Basaran E, Gökmen M, Kamasak ME, Shah M. Human semantic parsing for person re-identification. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA: IEEE; 2018. p. 1062–71. doi:10.1109/CVPR.2018.00117.
6. Liu Z, Huang Z, Xie D, Tian F, Li T. Person re-identification for suppressing background interference. J Comput-Aided Des Comput Graph. 2022;34(4):563–9. doi:10.3724/SP.J.1089.2022.18927.
7. Li S, Bak S, Carr P, Wang X. Diversity regularized spatiotemporal attention for video-based person re-identification. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA: IEEE; 2018. p. 369–78. doi:10.1109/CVPR.2018.00046.

8. Li M, Ji G. Research progress on video-based person re-identification. J Nanjing Normal Univ (Nat Sci Ed). 2020;43(2):120–30. doi:10.3969/j.issn.1001-4616.2020.02.019.
9. Sun Y, Zheng L, Deng W, Wang S. SVDNet for pedestrian retrieval. In: 2017 IEEE International Conference on Computer Vision (ICCV); 2017 Oct 22–29; Venice, Italy: IEEE; 2017. p. 3820–8. doi:10.1109/ICCV.2017.410.
10. Chen Y, Zhu X, Gong S. Person re-identification by deep learning multi-scale representations. In: 2017 IEEE International Conference on Computer Vision Workshops (ICCVW); 2017 Oct 22–29; Venice, Italy: IEEE; 2017. p. 2590–600. doi:10.1109/ICCVW.2017.304.
11. Xu SJ, Liu QY, Shi Y, Meng YB, Liu GH, Han JQ. Person re-identification based on diversified local attention network. J Electron Inf Technol. 2022;44(1):211–20 (In Chinese). doi:10.11999/JEIT201003.
12. Zhang L, Che J, Yang Q. Multi-granularity feature fusion for person re-identification. Chin J Liq Cryst Disp. 2020;35(6):555–63 (In Chinese). doi:10.3788/YJYXS20203506.0555.
13. Li D, Chen X, Zhang Z, Huang K. Learning deep context-aware features over body and latent parts for person re-identification. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA: IEEE; 2017. p. 7398–407. doi:10.1109/CVPR.2017.782.
14. Suh Y, Wang J, Tang S, Mei T, Lee KM. Part-aligned bilinear representations for person re-identification. In: Computer vision-ECCV 2018. Cham: Springer International Publishing; 2018. p. 418–37. doi:10.1007/978-3-030-01264-9_25.
15. Yang Q, Che J, Zhang L, Zhang YX. Person re-identification of GAN network hybrid coding. Chin J Liq Cryst Disp. 2021;36(2):334–42. doi:10.37188/CJLCD.2020-0167.
16. Lv J, Chen W, Li Q, Yang C. Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA: IEEE; 2018. p. 7948–56. doi:10.1109/CVPR.2018.00829.
17. Huang W, Li Y, Zhang K, Hou X, Xu J, Su R, et al. An efficient multi-scale focusing attention network for person re-identification. Appl Sci. 2021;11(5):2010. doi:10.3390/app11052010.
18. Song Y, Liu S, Yu S, Zhou S. Adaptive label allocation for unsupervised person re-identification. Electronics. 2022;11(5):763. doi:10.3390/electronics11050763.
19. Zhu Z, Chen S, Qi G, Li H, Gao X. Multi-granular inter-frame relation exploration and global residual embedding for video-based person re-identification. Signal Process Image Commun. 2025;132(6):117240. doi:10.1016/j.image.2024.117240.
20. Batool E, Gillani S, Naz S, Bukhari M, Maqsood M, Yeo SS, et al. POSNet: a hybrid deep learning model for efficient person re-identification. J Supercomput. 2023;79(12):13090–118. doi:10.1007/s11227-023-05169-4.
21. Wang J, Zhao Q, Jia D, Huang Z, Zhang M, Ren X. Spatial-temporal aware network for video-based person re-identification. Multimed Tools Appl. 2024;83(12):36355–73. doi:10.1007/s11042-023-16911-8.
22. Behera NKS, Sa PK, Bakshi S, Bilotti U. Explainable graph-attention based person re-identification in outdoor conditions. Multimed Tools Appl. 2023;483(107):210. doi:10.1007/s11042-023-16986-3.
23. Zheng L, Shen L, Tian L, Wang S, Wang J, Tian Q. Scalable person re-identification: a benchmark. In: 2015 IEEE International Conference on Computer Vision (ICCV); 2015 Dec 7–13; Santiago, Chile: IEEE; 2015. p. 1116–24. doi:10.1109/ICCV.2015.133.
24. Zheng Z, Zheng L, Yang Y. Unlabeled samples generated by GAN improve the person re-identification baseline *in vitro*. In: 2017 IEEE International Conference on Computer Vision (ICCV); 2017 Oct 22–29; Venice, Italy: IEEE; 2017. p. 3774–82. doi:10.1109/ICCV.2017.405.
25. Sun Y, Zheng L, Yang Y, Tian Q, Wang S. Beyond part models: person retrieval with refined part pooling (and A strong convolutional baseline). In: Computer vision-ECCV 2018. Cham: Springer International Publishing; 2018. p. 501–18. doi:10.1007/978-3-030-01225-0_30.
26. Lin Y, Zheng Z, Zhang H, Gao C, Yang Y. Bayesian query expansion for multi-camera person re-identification. Pattern Recognit Lett. 2020;130(10):284–92. doi:10.1016/j.patrec.2018.06.009.
27. Zhang L, Xiang T, Gong S. Learning a discriminative null space for person re-identification. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30; Las Vegas, NV, USA: IEEE; 2016. p. 1239–48. doi:10.1109/CVPR.2016.139.

28. Zhang XH. Improved person re-identification based on global feature. Comput Syst Appl. 2022;31(5):298–303 (In Chinese). doi:10.15888/j.cnki.csa.008477.
29. Sarfraz MS, Schumann A, Eberle A, Stiefelhagen R. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23. Salt Lake City, UT, USA: IEEE; 2018. p. 420–9. doi:10.1109/CVPR.2018.00051.
30. Su C, Li J, Zhang S, Xing J, Gao W, Tian Q. Pose-driven deep convolutional model for person re-identification. In: 2017 IEEE International Conference on Computer Vision (ICCV); 2017 Oct 22–29; Venice, Italy: IEEE; 2017. p. 3980–9. doi:10.1109/ICCV.2017.427.