Video Person Re-Identification with Frame Sampling–Random Erasure and Mutual Information–Temporal Weight Aggregation

Jiayue Li 1,2 and Yan Piao 1,*

1 Information and Communication Engineering, Electronics Information Engineering College, Changchun University of Science and Technology, Changchun 130022, China; lijiayue_87@126.com
2 High-Speed Railway Comprehensive Technical College, Jilin Railway Technology College, Jilin 132299, China
* Correspondence: piaoyan@cust.edu.cn

Abstract: Partial occlusion and background clutter in camera video surveillance affect the accuracy of video-based person re-identification (re-ID). To address these problems, we propose a person re-ID method based on random erasure of frame sampling and temporal weight aggregation of mutual information of partial and global features. First, for the case in which the target person is interfered or partially occluded, the frame sampling–random erasure (FSE) method is used for data enhancement to effectively alleviate the occlusion problem, improve the generalization ability of the model, and match persons more accurately. Second, to further improve the re-ID accuracy of video-based persons and learn more discriminative feature representations, we use a ResNet-50 network to extract global and partial features and fuse these features to obtain frame-level features. In the time dimension, based on a mutual information–temporal weight aggregation (MI–TWA) module, the partial features are added according to different weights and the global features are added according to equal weights and connected to output sequence features. The proposed method is extensively experimented on three public video datasets, MARS, DukeMTMC-VideoReID, and PRID-2011; the mean average precision (mAP) values are 82.4%, 94.1%, and 95.3% and Rank-1 values are 86.4%, 94.8%, and 95.2%, respectively.

Keywords: video person re-identification; mutual information–temporal weight aggregation (MI–TWA); frame sampling–random erasure (FSE); deep learning

1. Introduction

Person re-ID aims to retrieve the identities of specific persons captured by non-overlapping cameras in different environments [1]. Specifically, given a query image or video of a target person, the goal of person re-ID is to identify this specific person [2–4] from a multi-camera web gallery or video library in non-overlapping view domains. With the increased use of surveillance cameras, such as intelligent traffic systems, the wide application of multi-person tracking detection, and the rapid development of intelligent video analysis, the emphasis on public security monitoring and the popularization of intelligent transportation, person re-ID technology has become one of the most important areas in the field of video surveillance.

For image-based person re-ID, extracting effective pedestrian appearance features is an important step. However, the information contained in the image is limited. If the pedestrian information from a camera perspective is seriously missing, it will be difficult to find additional information for identification. Therefore, research on video-based person re-identification is particularly important. The purpose of video-based person re-ID [5–7] is to determine whether the person captured by video images from different cameras matches the target person. Compared with image-based person re-ID, more information is available, including information such as pedestrian appearance, pose, and motion. At the same time, it is closer to the actual application scenario because the pedestrians in the monitoring are also set forth in the form of video.
This article focuses on video-based pedestrian re-identification. With the widespread application of deep learning, Recurrent Neural Network (RNN) has also been applied to aggregate image-level features into sequence-level features for video-based person re-identification [8–11]. Some researchers also use optical flow to build spatiotemporal models with two-stream networks [12]. All these methods significantly improve the performance of video-based person re-identification. However, feature learning of deep learning requires a large number of samples to train the model, so it is easy to overfit on the dataset. Significant appearance variations due to camera viewpoint, background clutter, and especially, partial occlusion [13] can also occur when using this method. The performance of video-based person re-ID usually deteriorates severely under partial occlusion. This problem is difficult to solve because any part of a person may be occluded by other persons and environmental objects, such as bicycles and indicators. Attention mechanisms have been introduced into video-based person re-ID to handle partial occlusion [14]. They select discriminative frames from video sequences and generate informative video representations. While these methods have some tolerance for partial occlusion, discarding occluded frames is not ideal.

To address the occlusion problem of person re-ID, we use an improved ResNet-50 network to extract global and partial features and add frame-sampling–random erasure (FSE) and a mutual information–temporal weight aggregation (MI–TWA) module to obtain a more discriminative person feature representation. Not only can this method improve the training speed of the model, but it can also compensate for the lack of person image data and alleviate the problem of network overfitting.

In summary, our main contributions are as follows:

- A method using FSE is proposed for data enhancement to compensate for the lack of person image data and combine the length of frame sampling time to alleviate the problems of occlusion and noise in video frames.
- MI–TWA module is added to the improved ResNet-50 network [15], which is integrated with global features of equal weight to help reduce the impact of the occluded parts.
- Experimental results obtained using public benchmark datasets show that the proposed method based on FSE and MI–TWA can effectively solve the occlusion problem and improve the accuracy of video-based person re-ID.

2. Related Work

2.1. Image-Based Person Re-ID

Person re-ID is a challenging task that has been studied for many years. However, it still faces the same problems as other subfields of computer vision, including various poses, lighting, viewpoints, and occlusions. Person re-ID research has mainly focused on two sub-tasks, namely image-based [3,4,16–21] and video-based [22–26] person re-ID. Sun et al. [21] proposed a method that can accurately divide part-level features without resorting to pose estimation, which effectively improves the performance of re-ID. Zhang et al. [23] proposed a dense semantic contrast pedestrian re-identification method, which is also the first method to use fine-grained semantics to solve the problem of pedestrian image misalignment. In our proposed method, the frame-level image features are extracted according to the ResNet-50 and Part-based Convolutional Baseline (PCB) frameworks, respectively, to extract global features and partial features obtained by horizontal partitioning, so that they can complete subsequent temporal aggregation.

2.2. Video-Based Person Re-ID

Compared with image-based person re-ID, the samples in video-based person re-ID contain more frames and additional temporal information. Therefore, some existing methods have attempted to model additional temporal information to enhance video representations, and other methods have used 3D-CNNs [27,28] to explore spatiotemporal cues. For example, for spatial information, Li et al. [23] attempted to extract aligned spatial features.
across multiple images through different spatial attention modules, and Yang et al. [29] proposed a spatial-temporal GCN (STGCN) for mining spatial and temporal relations. Another class of methods [8,30] has utilized the additional information of optical flow and adopted a two-stream structure [31] for discriminative feature learning. However, optical flow only represents the local dynamics of adjacent frames and may introduce noise because of spatial asymmetry. The VRSTC [25] framework proposed by Hou et al. applied a generative adversarial network approach to the study of video-based person re-ID for the first time and recovered occluded local areas through upper and lower frame information. Using de-occluded video for feature learning can effectively solve occlusion problems. The random erasure method proposed by Zheng et al. [32] has also been applied in person re-ID. In person images, the problem of inaccurate matching caused by partial occlusion or incompleteness is solved by randomly selecting a rectangular area and erasing it. Unlike the previous methods, in this paper, we perform frame sampling-based random erasure in the image sequence input stage to reduce overfitting, improve the network generalization ability, and solve the occlusion problem in person re-ID. We strive to introduce the concept of mutual information into the temporal weight aggregation module and aggregate the extracted frame-level global and local features by assigning temporal weights according to the entropy value of mutual information to obtain better spatiotemporal features and reduce noise interference.

3. Methodology

3.1. Person Re-ID Network Based on Frame Sampling–Random Erasure and Mutual Information–Temporal Weight Aggregation

The overall structure of the proposed method is shown in Figure 1. First, the FSE operation is performed on the input video sequence, and the network model is used for feature extraction. The network model structure consists of a backbone network and three branch networks, namely, global feature branch, partial feature branch, and mutual information–temporal aggregation branch.

The backbone network uses ResNet-50 CNN and the last pooling layer and the fully connected layer of the network are removed. The global feature branch first extracts the features output by the fourth convolution block conv_4 in the ResNet-50 network and performs dimensional upgrade processing through 1 × 1 convolution. In order to preserve the differences in the characteristics of different pedestrians, BN (Batch Normalization) [33] is used for normalization to obtain the global feature \( f_c \). For the partial feature branch, we denote the feature map extracted from conv_4 and conv_5 of the ResNet-50 backbone as \( F \). \( F \) is horizontally partitioned into \( p \) parts, denoted as \( F_{pi} \), \( i = 1, \ldots, N \) [34]. To prevent overfitting through a bottleneck [15], each part of the feature map \( F_{pi} \) is processed by average pooling and 1 × 1 convolutional layers, obtaining a partial feature vector \( f_{cp_i} \).

Each partial feature \( f_{cp_i} \) is fed into a corresponding fully connected layer and a softmax layer is employed to predict the ID of each input image, denoted as \( f_{cp} \).

Based on the feature map channel attention mechanism, we designed a feature fusion network [15]. The network structure is shown in Figure 2. The global and partial features are inputs. After cascading, the bottleneck mechanism is used to build a channel attention mechanism to promote interactions between different features and information exchange. At the same time, the residual structure is used to accelerate and stabilize convergence so that, finally, the frame-level feature \( f_{c1} \) is obtained.
Figure 1. Architecture of our method. The FSE-processed video sequence is used as the input of the network. The ResNet-50 is used to extract the global feature $f_{c}\mathbf{q}$ and the partial feature $f_{c}\mathbf{p}$. $f_{c}\mathbf{q}$ and $f_{c}\mathbf{p}$ are fused to obtain the frame-level image feature $f_{c}\mathbf{t}$. $f_{c}\mathbf{p}$ and $f_{c}\mathbf{q}$ are aggregated by the temporal weights. The temporal weight $W_{t}$ of $f_{c}\mathbf{p}$ is calculated by mutual information entropy value. The temporal weight of $f_{c}\mathbf{q}$ is calculated by equal weighting. $f_{c}\mathbf{p}'$ and $f_{c}\mathbf{q}'$ are concatenated as the final feature representation of this sequence. By combining $L_{triplet}$ and $L_{softmax}$, the network model is trained.

Figure 2. Feature fusion network.

The mutual information–temporal aggregation branch combines global features with equal weights, aggregates partial features with different time weights according to the entropy value of mutual information, and obtains sequence features by connecting the weighted global features to the partial features; finally, the triple loss function [35] and the softmax loss function [27,36] are combined to train the entire network.

3.2. Data Enhancement of Frame Sampling–Random Erasure

As the length of the video sequence varies greatly in the model, the visual information of the entire video is used to solve this problem. For the input video, the entire video is evenly divided into $T$-consecutive person trajectory sequences. During training, each person’s trajectory sequence is randomly sampled in any frame, and then a sampled frame with some temporal information is taken as input.

For the situation where the target pedestrian is disturbed or partially occluded, to solve the occlusion problem and improve the generalization ability of the model, random erasure is performed after frame sampling to occlude parts of the image at random positions.
This method can generate enhanced images with various degrees of occlusion, making the model more robust to noise and occlusion. For a video sampling frame in a batch, we randomly selected a rectangular area $f_e$ from the image frame $S = W \times H$ for pixel erasure. The height and width ratios of the random erasure area are $r_h$ and $r_w$, respectively. The random erase area is $f_e = W_h \times H_e$, $W_e = W \times r_w$, and $H_e = H \times r_h$. A point $P(x_e, y_e)$ is randomly selected in the image as $x_e + W_e \leq W$ and $y_e + H_e \leq H$, then the area $I_e = (x_e, y_e, x_e + W_e, y_e + H_e)$ is selected as the random rectangular area to be erased. Otherwise, we continue searching for the point $P$ until the required area $I_e$ is selected, as well as training images with different degrees of occlusion are generated to enhance the data. Detailed flow of the FSE data augmentation method is shown in Figure 3.

![Figure 3. Pipeline of frame sampling–random erasure.](image)

### 3.3. Mutual Information–Temporal Weight Aggregation

Next, we need to merge the frame-level features from the sampled sequence in the temporal dimension. Because of changes in human poses, occlusion, and viewpoint, not all frames have the same amount of information. We are more concerned with frames that provide explicit person feature information, so we calculate the weight $W_{it}$ of the $t$-th frame in the $i$-th part by the entropy value of the mutual information.

Our method sets the local feature horizontal part $N$, $N = 6$. Useful information related to the non-random erasure part in the partial feature and the person in the global feature is defined as mutual information $MI$ (Mutual Information), which is denoted as follows:

$$MI_i(fc) = p(fc^i) \log \frac{p(fc^i, fc)}{p(fc^i)p(fc)}$$

where $p(fc^i)$ denotes the probability that a local feature appears in the entire global feature; $p(fc)$ denotes the probability of the person part in the entire global feature; $p(fc^i, fc)$ denotes the ratio of the number of the person part in the global feature that contains partial features and the number of the entire global feature.

The entropy value of the amount of information is denoted as follows:

$$E(fc^i) = MI_i(fc) \sum_{i=1}^{N} p(fc^i) \log(p(fc^i)),$$

then, the mutual information entropy value of the $t$ frame is normalized to obtain the time weight as follows:

$$W_{it} = \frac{1 - E(fc^i)}{\sum_{i=1}^{N} \left[ 1 - E(fc^i) \right]}.$$
the partial features are aggregated with and temporal weights, which are calculated as follows:

$$f_{c}^{p} = \sum_{t=1}^{T} W_{t} f_{c}^{p}_{t},$$  \quad (4)

the global features are merged with equal weights $1/t$, which is calculated as follows:

$$f_{c}^{q} = \sum_{t=1}^{T} \frac{1}{t} f_{c}^{q}_{t}.$$  \quad (5)

Finally, the partial features $f_{c}^{p}$ and global features $f_{c}^{q}$ are connected to represent sequence features. In the train stage, combination of cross entropy loss and batch triplet loss was used to train the network. In the test stage, Euclidean distance of the connected features for $L_2$ normalization was used to evaluate the similarity of the video sequences.

3.4. Loss Function

The same loss function in the global feature branch and the partial feature branch, $L_{\text{triplet}}$ [35] and $L_{\text{softmax}}$ [27,36], are set. $L_{\text{softmax}}$ is denoted as follows:

$$L_{\text{softmax}} = -\frac{1}{PK} \sum_{i=1}^{P} \sum_{n=1}^{K} p_{i,n} \log q_{i,n},$$  \quad (6)

where $p_{i,n}$ and $q_{i,n}$ are the true and predicted identities of the samples.

$$L_{\text{triplet}} = -\sum_{i=1}^{P} \sum_{a=1}^{K} \alpha + \max_{p=1\ldots K} \left\| f_{a}^{(i)} - f_{p}^{(i)} \right\|_2 - \min_{n=1\ldots K, j=1\ldots P} \left\| f_{a}^{(i)} - f_{n}^{(j)} \right\|_2 +$$  \quad (7)

In Formula (7), $f_{a}$, $f_{p}$, $f_{n}$ are features extracted from target samples, positive samples, and negative samples. Positive samples are persons with the same identity as the target sample; negative samples are persons with different identities from the target samples; $\alpha$ is the hyperparameter used to control the distance within the sample; $P$ is the number of pedestrians extracted when training each batch; $K$ corresponds to the number of images extracted for each person during training. By combining $L_{\text{triplet}}$ and $L_{\text{softmax}}$, the network model is optimized, guiding the network to explore according to the features of the discriminative power; $L$ is denoted as follows:

$$L = L_{\text{triplet}} + L_{\text{softmax}}.$$  \quad (8)

4. Experiments

The experimental environment in this paper is based on Ubuntu 20.04, Cuda10.2 and Pytorch 1.8.1, and torchvision 0.9.0 deep-learning framework. The hardware configuration includes three GPUs, GTX 3070Ti (8 GB video memory), and the programming language Python 3.8. To demonstrate the effectiveness of our method on the problem of person re-ID, we evaluate it on three benchmark datasets—MARS, DukeMTMC-VideoReID, and PRID-2011, and compare it with state-of-the-art methods.

4.1. Datasets and Evaluation Protocols

Datasets. MARS [37] is one of the largest video person re-ID datasets to date. The MARS dataset includes 1261 persons captured by six cameras and each person has at least one sequence of person trajectories from two cameras. All bounding boxes and trajectories
of MARS are automatically generated, so they are large in scale and difficult to match at the same time, which has the characteristics of false detection interference.

The PRID-2011 [38] dataset provides video clips of multiple people under two different static surveillance cameras. Cameras A and B have 385 people and 749 people, respectively, and 200 people appear in both the A and B perspectives. The length of the video sequences varies from 5 to 675 frames, with an average of 100 frames. The dataset is collected in uncrowded outdoor scenes with relatively simple and clean backgrounds with little occlusion.

The DukeMTMC-VideoReID [39] dataset is a subset of the DukeMTMC dataset for video-based person re-ID. It consists of images collected by eight cameras with different viewing angles at the Duke University campus, with 36,441 manually detected images of 1812 persons. Because the dataset is manually annotated, for each identity mark there is only one video clip. The dataset includes 702 identities for training, 702 identities for testing, and 408 identities as distractors. Table 1 presents the statistics of the three video datasets.

| Dataset                  | #ID   | #Seq  | #Train | #Test | #Cam | #Label                       |
|--------------------------|-------|-------|--------|-------|------|------------------------------|
| PRID-2011                | 200   | 400   | 89     | 89    | 2    | Manual annotation            |
| MARS                     | 1261  | 20478 | 625    | 636   | 6    | Deformable Part Models (DPM)  |
| DukeMTMC-VideoReID       | 1812  | 4832  | 702    | 702   | 8    | Manual annotation            |

**Evaluation Metrics.** The metrics of cumulative matching characteristic (CMC) and mean average precision (mAP) are used for evaluation [23,37,40,41]. CMC is based on the retrieval capability of the algorithm to find the correct identity within the top-k ranked matches. We report Rank-1, Rank-5, and Rank-10 CMC accuracies. The mAP metric is used to evaluate algorithms in multi-shot re-identification settings where multiple instances of the same identities are present in the gallery.

**Experimental parameter settings.** This experiment uses ResNet-50 pre-trained on ImageNet as the backbone network. For network training, we use the Adam optimizer, which iteratively updates the neural network weights based on the training data. First, we randomly select \(T = 8\) frames from the input video sequence and then randomly select \(P = 8\) identities to sample each mini batch. We randomly sample \(K = 4\) videos for each identity from the training set sequence and then perform random erasure, setting the erasure height ratio to 1/3. We also resize images to 128 \(\times\) 256. To make the objective function converge and optimize the network model, the initial value of the learning rate is set to 0.01, and it is reduced to 1/10 of the previous value every 20 cycles until the learning rate reaches 0.0001. The margin parameter of the triple loss function is set to 0.3.

### 4.2. Experimental Evaluation

#### 4.2.1. FSE and MI–TWA Partial Features Improve Performance

To verify the effectiveness of the proposed method, three datasets of video person re-ID are tested and analyzed. The experimental results obtained by using the baseline method, the baseline + FSE method, and the baseline + MI–TWA are presented in Table 2.

| Dataset                  | mAP | Rank-1 | Rank-5 | Rank-10 | mAP | Rank-1 | Rank-5 | Rank-10 | mAP | Rank-1 | Rank-5 | Rank-10 |
|--------------------------|-----|--------|--------|---------|-----|--------|--------|---------|-----|--------|--------|---------|
| MARS                     | 73.8| 80.2   | 89.1   | 92.3    | 84.2| 86.8   | 88.2   | 91.9    | 86.3| 85.1   | 91.7   | 95.2    |
| DukeMTMC-VideoReID       | 76.3| 82.8   | 90.5   | 92.9    | 88.4| 90.6   | 91.6   | 93.3    | 89.2| 87.3   | 95.3   | 97.6    |
| PRID-2011                | 80.8| 84.5   | 91.6   | 94.2    | 92.2| 92.6   | 95.5   | 96.3    | 92.5| 90.7   | 97.2   | 98.2    |
The baseline method uses ResNet-50 as the backbone network. After global average pooling, the feature vector of the video segment dimension is 2048 and it only contains the basic network structure of the global feature branch and the softmax cross-entropy loss function.

From Table 2, we observe the following effects of adding FSE to the baseline method: On the MARS dataset, Rank-1 increases from 80.2% to 82.8% and mAP increases from 73.8% to 76.3%. On the DukeMTMC-VideoReID dataset, Rank-1 increases from 86.8% to 90.6% and mAP increases from 84.2% to 88.4%. On the PRID-2011 dataset, Rank-1 increases from 85.1% to 87.3% and mAP increases from 86.3% to 89.2%. These results show that FSE can prevent overfitting and effectively improve recognition performance.

The effect of adding partial feature temporal weight aggregation using mutual information to the baseline method is as follows: On the MARS dataset, Rank-1 increases from 80.2% to 84.5% and mAP increases from 73.8% to 80.8%. On the DukeMTMC-VideoReID dataset, Rank-1 increases from 86.8% to 92.6% and mAP increases from 84.2% to 92.2%. On the PRID-2011 dataset, Rank-1 increases from 85.1% to 90.7% and mAP increases from 86.3% to 92.5%. The experimental results show that in the case of changes and occlusions, partial feature temporal weight aggregation can effectively refine temporal features and increase the distinguishability of feature representation by setting different temporal weights according to the different amounts of information provided by the partial features of all frames.

### 4.2.2. Performance Improvement of Partial Feature Method with Joint FSE and MI–TWA

From Table 3, we observe that when the FSE method is used with Baseline + MI–TWA, the accuracy improves significantly. Baseline + MI–TWA+FSE outperforms Baseline + MI–TWA by 4.5% (95.2%/90.7%) Rank-1, 2.2% (94.8%/92.6%) Rank-1, and 1.9% (86.4%/84.5%) Rank-1, respectively, and by 2.8% (95.3%/92.5%) mAP, 1.9% (94.1%/92.2%) mAP, and 1.6% (82.4%/80.8%) mAP, respectively, on the PRID-2011, DukeMTMC-VideoReID, and MARS datasets. The experimental results fully demonstrate that the proposed method can effectively improve the performance of person re-ID.

Table 3. Comparison between Baseline + MI–TWA+FSE and Baseline + MI–TWA models. The Rank-i CMC accuracies and mAP scores are reported.

| Dataset | PRID-2011 | DukeMTMC-VideoReID | MARS | PRID-2011 | DukeMTMC-VideoReID | MARS |
|---------|-----------|-------------------|------|-----------|-------------------|------|
| Rank-1  | 90.7      | 92.6              | 84.5 | 95.2      | 94.8              | 86.4 |
| Rank-5  | 97.2      | 95.5              | 91.6 | 97.3      | 96.1              | 92.3 |
| Rank-10 | 98.2      | 96.3              | 94.2 | 98.5      | 96.6              | 94.7 |
| mAP     | 92.5      | 92.2              | 80.8 | 95.3      | 94.1              | 82.4 |

### 4.2.3. Influence of Video Sequence Length T and Random Erasure Height Ratio on Performance

Table 4 presents the results of comparing the performances of video sequences with different lengths. Besides changing the length $T$ of the video sequence, other parameters remain unchanged. $T = 1$ is a model for a single image that does not use temporal features. From Table 4, we observe that an increase in the sequence length $T$ results in an improvement in mAP and rank accuracy scores. Compared with $T = 4$, $T = 8$ results in a slight change in both mAP and rank indicators. However, both $T$ values have high indicators. Therefore, choosing $T = 4$ in the experiment can provide good results.
Table 4. Performance comparison of different video sequence lengths.

| T   | Rank-1 | Rank-5 | Rank-10 | mAP  | Rank-1 | Rank-5 | Rank-10 | mAP  | Rank-1 | Rank-5 | Rank-10 | mAP  |
|-----|--------|--------|---------|------|--------|--------|---------|------|--------|--------|---------|------|
| 1   | 88.7   | 89.1   | 92.3    | 83.7 | 79.0   | 86.9   | 90.2    | 71.1 | 81.7   | 86.6   | 89.6    | 81.2 |
| 2   | 92.3   | 94.8   | 96.4    | 91.8 | 91.8   | 92.9   | 90.1    | 76.7 | 88.2   | 92.6   | 94.7    | 88.5 |
| 4   | 95.2   | 97.3   | 98.5    | 95.3 | 95.3   | 96.4   | 92.3    | 94.7 | 94.8   | 96.1   | 96.6    | 94.1 |
| 8   | 95.0   | 97.2   | 98.2    | 95.0 | 95.0   | 96.0   | 92.2    | 94.4 | 92.6   | 96.0   | 96.3    | 92.8 |

From Table 5, we observe that when the video length is $T = 4$, different erasure height ratios have different effects on the network performance. When the ratio is 1/2, the network convergence is relatively poor and the index is relatively low. When the ratio is 1/3, the network has a certain advantage in mAP and the Rank-1, Rank-5, and Rank-10 values achieve the best results. However, when the ratio is 1/2, the erasure area is too large, obscuring a significant amount of information with discriminative rows, thus, resulting in poor network performance.

Table 5. Comparison of the impact of different erasure height ratios on network performance.

| Ratio | MARS | DukeMTMC-VideoReID |
|-------|------|---------------------|
| 1/2   | 82.3 | 91.4 93.8 75.6 88.7 93.8 94.7 89.2 |
| 1/3   | 86.4 | 92.3 94.7 82.4 94.8 96.1 96.6 94.1 |

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

We compare the proposed method with state-of-the-art video-based re-ID methods, including MARS [37], Temporal Attention Model + Spatial Recurrent Model (TAM + SRM), Sequential Decision Making (SDM), VRSTC, GLTR, Attribute-Driven Feature Disentangling and Temporal Aggregation (ADFD-TA), and other methods. The results are summarized in Table 6.

Table 6. Comparisons of our proposed approach to the state-of-the-art methods.

| Method       | MARS | DukeMTMC-VideoReID | PRID-2011 |
|--------------|------|---------------------|-----------|
|              | Rank-1 | mAP  | Rank-1 | mAP  | Rank-1 | mAP  | Rank-1 | mAP  |
| MARS [37]    | 68.3 49.3 | -    | -      | -    | 77.3  | -    |
| TAM + SRM [9]| 70.6 50.7 | -    | -      | -    | 79.4  | 94.4 |
| SDM [42]     | 71.2  | -    | -      | -    | 85.2  | -    |
| ADFD-TA [43]| 82.6 71.2 | -    | -      | -    | 93.9  | -    |
| GLTR [33]    | 87.0 78.5 | 96.3 93.7 | 95.5  | -    |
| VRSTC [25]   | 88.5 82.3 | 95.0 93.5 | -    | -    |
| MI–TWA + FSE(Ours) | 86.4 82.4 | 94.8 94.1 | 95.2 95.3 |

From Table 6, on the PRID-2011 small video dataset, our method can achieve 95.2% in Rank-1 and 95.3% in mAP. On the newer video dataset, DukeMTMC-VideoReID, our method can achieve 94.8% in Rank-1 and 94.1% in mAP. On the more challenging video dataset, MARS, our method can achieve 86.4% in Rank-1 and 82.4% in mAP. Compared with GLTR [33] and VRSTC [25], our method is lower in Rank-1, but higher in mAP than GLTR [33] and VRSTC [25]; Rank-1 has a series of random factors, while mAP is a more comprehensive evaluation metric, which can indicate better performance. We observe that the method of joint FSE and MI–TWA proposed in this paper is better than
the current state-of-the-art methods in comprehensive performance. The proposed method effectively enriches person feature information and further improves the video person re-ID recognition rate.

To further verify the generalization ability of the network model, cross-domain testing experiments are performed. The results are shown in Table 7. In Table 7, M and D represent MARS and DukeMTMC-VideoReID, respectively; the letters on the left and right of the arrow ("→") represent the training set and the test set, respectively. We find that Rank-1 and mAP of the proposed model can only reach 58.6% and 35.7%. Analysis of the results shows that the training set and test set in the person re-ID dataset are disjoint and the features of each dataset are quite different. For example, in the MARS dataset, persons mostly wear summer clothes, such as short sleeves and shorts; however, the pedestrians in DukeMTMC-VideoReID are wearing jackets and trousers and the colors are relatively dull, thus, leading to poor cross-domain performance of the model. The experimental results show that the generalization ability of the proposed model has been improved to a certain extent, but the overall cross-domain recognition ability needs to be further improved.

Table 7. Performance of our models on the cross-domain condition.

| Model         | M→D  | D→M  |
|---------------|------|------|
| Rank-1        | 48.2 | 58.6 |
| mAP           | 31.5 | 35.7 |

5. Conclusions

In this paper, the temporal modelling method based on video person re-ID is improved and the weight distribution of the partial feature information entropy values on the temporal information of persons is refined, which significantly improves the accuracy of video-based person re-ID. Moreover, frame sampling is adopted. A data augmentation method of random erasure enriches the representation of video person features. We conduct evaluation experiments on the small PRID-2011 dataset and the large DukeMTMC-VideoReID video dataset. A series of comparative experiments are also performed on the large, highly representative MARS video dataset. Experimental results on the three video datasets show that the proposed method of joint FSE and MI–TWA can effectively extract discriminative person feature representations, solve person occlusion problems, and achieve accurate person recognition. It is superior to many existing video-based person re-ID methods in terms of degree and efficiency; however, its cross-domain recognition rate needs to be improved. The main task of future work is to improve the recognition rate of cross-domain testing and combine the proposed method with a target detection or tracking algorithm and apply it to the actual multi-camera monitoring environment to achieve accurate recognition and continuous and stable tracking of target persons.

Author Contributions: Conceptualization, Y.P. and J.L.; methodology, Y.P. and J.L.; software, J.L.; validation, J.L.; investigation, J.L.; writing—original draft preparation, J.L.; writing—review and editing, Y.P. and J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.
References
1. Song, C.; Huang, Y.; OuYang, W.; Wang, L. Mask-guided contrastive attention model for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–21 June 2018; pp. 1179–1188.
2. Chen, T.L.; Ding, S.J.; Xie, J.Y.; Yuan, Y.; Chen, W.; Yang, Y.; Ren, Z.; Wang, Z. ABD-Net: Attentive but diverse person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–21 June 2018; pp. 2285–2294.
3. 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 Proceedings of the 2018 European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 480–496.
4. Zheng, M.; Karanam, S.; Wu, Z.; Radke, R.J. Re-identification with consistent attentive siamese networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–21 June 2019; pp. 5735–5744.
5. Fu, Y.; Wang, X.Y.; Wei, Y.C.; Huang, T. STA: Spatial-Temporal Attention for Large-Scale Video-Based Person Re-Identification. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI), Honolulu, HI, USA, 27 January–1 February 2019; pp. 8287–8294.
6. Liu, X.H.; Zhang, P.P.; Yu, C.Y.; Lu, H.C.; Qian, X.S.; Yang, X.Y. A video is worth three views: Trigeminal transformers for video-based person reidentification. arXiv 2021, arXiv:2104.01745.
7. Li, S.Z.; Yu, H.M.; Hu, H.J. Appearance and Motion Enhancement for Video-Based Person Re-Identification. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI), New York, NY, USA, 7–12 February 2020; pp. 11394–11401.
8. McLaughlin, N.; Rincon, J.M.; Miller, P. Recurrent Convolutional Network for Video-Based Person Re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1325–1334.
9. Zhou, Z.; Huang, Y.; Wang, W.; Wang, L.; Tan, T.N. See the Forest for the Trees: Joint Spatial and Temporal Recurrent Neural Networks for Video-Based Person Re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4747–4756.
10. Dai, J.; Zhang, P.P.; Wang, D.; Lu, H.C.; Wang, H.Y. Video Person Re-Identification by Temporal Residual Learning. IEEE Trans. Image Process. 2019, 28, 1366–1377. [CrossRef] [PubMed]
11. Wu, Z.; Wang, X.; Jiang, Y.G.; Ye, H.; Xue, X. Modeling spatial-temporal clues in a hybrid deep learning framework for video classification. In Proceedings of the ACM Multimedia (ACM), Sydney, Australia, 11–16 October 2015; pp. 461–470.
12. Yu, Z.; Li, T.; Yu, N.; Gong, X. Three-stream convolutional networks for video-based person re-identification. arXiv 2017, arXiv:1712.01652.
13. He, L.X.; Liang, J.; Li, H.Q.; Sun, Z.N. Deep spatial feature reconstruction for partial person reidentification: Alignment-free approach. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–21 June 2018.
14. Zhang, S.S.; Yang, J.; Schiele, B. Occluded pedestrian detection through guided attention in cnns. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–21 June 2018.
15. He, K.M.; Zhang, X.Y.; Ren, S.Q.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
16. Ge, Y.; Gu, X.; Chen, M.; Wang, H.; Yang, D. Deep multi-metric learning for person re-identification. In Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), San Diego, CA, USA, 23–27 July 2018; pp. 1–6.
17. Zhang, Z.; Lan, C.; Zeng, W.; Chen, Z. Densely semantically aligned person reidentification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–21 June 2019; pp. 667–676.
18. Chen, D.; Xu, D.; Li, H.; Sebe, N.; Wang, X. Group consistent similarity learning via deep crf for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–21 June 2018; pp. 8649–8658.
19. Liu, F.; Zhang, L. View confusion feature learning for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–21 June 2019; pp. 6639–6648.
20. Hou, R.; Ma, B.; Chang, H.; Gu, X.; Shan, S.; Chen, X. Interaction-and-aggregation network for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–21 June 2019; pp. 9317–9326.
21. Gu, X.; Ma, B.; Chang, H.; Shan, S.; Chen, X. Temporal knowledge propagation for image-to-video person re-identification. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–3 November 2019; pp. 9647–9656.
22. Liu, Y.; Yan, J.; OuYang, W. Quality aware network for set to set recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4694–4703.
23. Li, S.; Bak, S.; Carr, P.; He, T.C.; Wang, X. Diversity regularized spatiotemporal attention for video-based person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–21 June 2018; pp. 369–378.
24. Xu, S.; Cheng, Y.; Gu, K.; Yang, Y.; Chang, S.; Zhou, P. Jointly attentive spatial temporal pooling networks for video-based person re-identification. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 4743–4752.

25. Hou, R.; Ma, B.; Chang, H.; Gu, X.; Shan, S.; Chen, X. VRSTC: Occlusion-free video person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–21 June 2019; pp. 7183–7192.

26. Gu, X.; Ma, B.; Chang, H.; Zhang, H.; Chen, X. Appearance-Preserving 3D Convolution for Video-Based Person Re-Identification. In Proceedings of the 2020 European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; pp. 228–243.

27. Liao, X.Y.; He, L.X.; Yang, Z.W.; Zhang, C. Video-Based Person Re-identification via 3D Convolutional Networks and Non-local Attention. In Proceedings of the 14th Asian Conference on Computer Vision (ACCV), Perth, Australia, 2–6 December 2018; pp. 620–634.

28. Li, J.N.; Zhang, S.L.; Huang, T.J. Multiscale 3d convolution network for video-based person identification. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI), Honolulu, HI, USA, 27 January–1 February 2019.

29. Yang, J.R.; Zheng, W.S.; Yang, Q.Z.; Chen, Y.C.; Tian, Q. Spatial-Temporal Graph Convolutional Network for Video-Based Person Re-Identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 3289–3299.

30. Chung, D.; Tahboub, K.; Edward, J.D. A two stream siamese convolutional neural network for person re-identification. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.

31. Karen, S.Y.; Andrew, Z. Two-stream convolutional networks for action recognition in videos. In Proceedings of the Conference on Neural Information Processing Systems(NIPS), Montreal, QC, Canada, 8–13 December 2014.

32. Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; Yang, Y. Random erasing data augmentation. arXiv 2017, arXiv:1708.04896. [CrossRef]

33. Sergey, I.; Christian, S. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In Proceedings of the International Conference on Machine Learning (ICML), Lille, France, 6–11 July 2015; pp. 448–456.

34. Miao, J.X.; Wu, Y.; Liu, P.; Ding, Y.H.; Yang, Y. Pose-Guided Feature Alignment for Occluded Person Re-Identification. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–3 November 2019; pp. 542–551.

35. Hermans, A.; Beyer, L.; Leibe, B. In defense of the triplet loss for person reidentification. arXiv 2017, arXiv:1703.07737.

36. Subramaniam, A.; Nambiar, A.; Mittal, A. Co-segmentation inspired attention networks for video-based person re-identification. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–3 November 2019.

37. Zheng, L.; Bie, Z.; Sun, Y.F.; Wang, J.D.; Su, C.; Wang, S.J.; Tian, Q. Mars: A video benchmark for large-scale person re-identification. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 10–16 October 2016; pp. 868–884.

38. Hirzer, M.; Beleznai, C.; Roth, P.M.; Bischof, H. Person re-identification by descriptive and discriminative classification. In Proceedings of the Scandinavian conference on Image analysis, Ystad, Sweden, 13–15 May 2011; pp. 91–102.

39. Suh, Y.M.; Wang, J.D.; Tang, S.Y.; Tao, M.; Lee, K.M. Part-aligned bilinear representations for person re-identification. In Proceedings of the 2018 European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 402–419.

40. Zhang, J.; Wang, N.; Zhang, L. Multi-shot pedestrian reidentification via sequential decision making. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–21 June 2018; pp. 5177–5186.

41. Liu, H.; Jie, Z.Q.; Jayashree, K.; Qi, M.B.; Jiang, J.G.; Yan, S.C.; Feng, J.S. Video-based person re-identification with accumulative motion context. IEEE Trans. Circuits Syst. Video Technol. 2018, 28, 2788–2802. [CrossRef]

42. Zhao, Y.; Shen, X.; Jin, Z.; Lu, H.; Hua, X.S. Attribute-driven feature disentangling and temporal aggregation for video person reidentification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–21 June 2019; pp. 4908–4917.