Temporal Aggregation with Clip-level Attention for Video-based Person Re-identification

Mengliu Li[1]*, Han Xu[2]*, Jinjun Wang[1]†, Wenpeng Li[2] and Yongli Sun[2]
[1]Institute of Artificial Intelligence and Robotics, Xi'an Jiao Tong University
[2]Deep North Inc., Xi'an, Shaanxi, China
[1]{meng1996liu@stu, jinjun@mail}.xjtu.edu.cn, [2]{hxu, wpli, ylsun}@deepnorth.cn

Abstract

Video-based person re-identification (Re-ID) methods can extract richer features than image-based ones from short video clips. The existing methods usually apply simple strategies, such as average/max pooling, to obtain the tracklet-level features, which has been proved hard to aggregate the information from all video frames. In this paper, we propose a simple yet effective Temporal Aggregation with Clip-level Attention Network (TACAN) to solve the temporal aggregation problem in a hierarchal way. Specifically, a tracklet is firstly broken into different numbers of clips, through a two-stage temporal aggregation network we can get the tracklet-level feature representation. A novel min-max loss is introduced to learn both a clip-level attention extractor and a clip-level feature representer in the training process. Afterwards, the resulting clip-level weights are further taken to average the clip-level features, which can generate a robust tracklet-level feature representation at the testing stage. Experimental results on four benchmark datasets, including the MARS, iLIDS-VID, PRID-2011 and DukeMTMC-VideoReID, show that our TACAN has achieved significant improvements as compared with the state-of-the-art approaches.

1. Introduction

As one of the most popular yet challenging computer vision tasks, person re-identification (Re-ID) aims at tracking pedestrians among multiple non-overlapping camera views. On one hand, person Re-ID is able to save a significant amount of human labour in exhaustively searching for the person of interest amongst a large set of video tracklets collected from multiple cameras. On the other hand, person Re-ID is also a very challenging problem because the same pedestrian's appearance changes drastically across the disjoint camera views, due to the changes in various covariate factors such as posture, background and illumination.

In the past few years, image-based person Re-ID has made great progress [21, 23, 30]. A typical image-based person Re-ID model usually takes a single cropped full-body image as input, and a pre-defined distance metric is utilized to measure the similarity between every pair of cropped images. In this way, a person is re-identified if the corresponding similarity is above a defined threshold [8]. To further improve the model's capacity, some researchers attempted part-based methods to make full use of body shape information [21]. For example, slicing images into several parts turned out to be a useful solution [23] in learning discriminative features for person Re-ID. Although the image-based person Re-ID methods have achieved considerable accuracy on public datasets, including the Market-1501 [28], DukeMTMC-reID [29], and CUHK03 [12], when applied to real-world data, the accuracy is still limited by various factors such as background, occlusion and posture variations. Besides, some researchers have also attempted to detect the foreground object from an image to focus features mostly on the salient parts [30], which on the other hand is also very challenging due to the difficulty in learning the attention from single images. To deal with the issue of image quality, some other researchers have also investigated the generative adversarial networks (GAN) [4] based method to reproduce frame images with less noise. However, these methods still suffer from the limited information available from single images.

Hence, video-based person Re-ID is attracting increasing popularity because videos usually contain richer information than single images. Besides, the video-based person Re-ID datasets are easily available, which makes it possible to overcome this problem. For example, the MARS [13] dataset is released as an extension of Market-1501, where the original Market-1501 dataset becomes a subset of MARS and images are carefully selected from the corresponding video tracklets. In fact, video tracklets are

*These authors contributed equally to this work and authorship is in alphabetical order.
†Corresponding author, jinjun@mail.xjtu.edu.cn

more readily available from many real-world surveillance systems. The challenges of video-based person Re-ID include how to select the most suitable frames to apply image-based models, how to aggregate information from multiple frames, and how to obtain a robust feature representation for person association. Extensive experimental results have shown that simple strategies, such as average/max pooling, to aggregate the consecutive frame-level features are not always effective in practice. What's worse, some existing methods directly extend the image-based ones [15, 20] to solve the video-based person Re-ID problem, where image features are extracted and then combined over tracklet. We argue that the video-based person Re-ID methods should focus on strategies of how to deal with consecutive frames and a well-designed strategy should also take advantage of rich temporal information of video tracklets that do not exist in single images.

In this paper, we propose a simple yet effective Temporal Aggregation with Clip-level Attention Network (TACAN) to solve the temporal aggregation problem in a hierarchal way. In the training process, we firstly break each tracklet into different numbers of clips, and then a novel min-max loss is introduced to learn both a clip-level attention extractor and a clip-level feature representer. At the testing stage, the resulting clip-level weights are further taken to average the clip-level features, which can generate a robust tracklet-level feature representation for video-based person Re-ID. In summary, our contributions in this paper are two folds: 1) We propose to learn the clip-level attention and propose a two-stage temporal aggregation network TACAN to effectively consider more temporal information when aggregating features over frames and clips; 2) We adopt the min-max loss into our framework which can noticeably improve the training efficiency. Experimental results on four benchmark datasets, including the MARS, iLIDS-VID, PRID-2011 and DukeMTMC-VideoReID, show that our TACAN has achieved significant improvements as compared with the state-of-the-art approaches.

2. Related Works

In this section, some related works on person Re-ID are listed, including image-based Re-ID methods, video-based Re-ID methods, and metric learning, to show massive achievements in these years on person Re-ID tasks.

2.1. Image-based Person Re-ID

Recent works on image-based person Re-ID is mainly focusing on two directions: image feature representation and metric learning. In the image feature representation task, normally, image-based Re-ID models have two parts in the pipeline. The first part is a base model for preliminary features and the other part is a fine-grained feature extractor. Since the rapid development of Convolutional Neural Network (CNN), it has become the mainstream using the image classification models as the feature extractor base model for person Re-ID. Many CNN models that pre-trained on the ImageNet dataset can be transferred to the person Re-ID tasks. ResNet [5] and SENet [10] are widely-used networks that perform well on ImageNet dataset. Based on the idea of residual learning, ResNet can speed up the training of neural networks, and the network can still learn representative features while increasing the depth. The 50-layer ResNet, which we called ResNet50, is widely used as the base model for many transfer learning tasks. SENet is on the basis of ResNet that squeezes and excites channel-wise feature responses by explicitly modeling inter-dependencies between channels. Generally speaking, SENet performs better than ResNet on image classification tasks.

Based on the development of CNN networks, many effective slice-and-segmentation-based person Re-ID methods have been proposed. Part-based Convolutional Baseline (PCB) [21] employed a simple uniform partition strategy and assembled part-informed features into a convolutional descriptor. Multiple Granularity Network (MGN) [23] divided the original images into 2 and 3 stripes and aggregated global and local features to obtain the final feature representation. Foreground attention neural network (FANN) [30] learned a discriminative feature representation through enhance the positive side of the foreground and weaken the negative side of the background.

2.2. Video-based Person Re-ID

Compared to former image-based person Re-ID, video-based Re-ID can take more advantages from temporal information. Since the works of extracting image features have many achievements, how to effectively aggregate multi-frame image features becomes the key problem of research. Many temporal modeling methods were proposed, such as Recurrent Neural Network (RNN), temporal attention, and 3D CNN [14]. It has been proven that Temporal attention models have the best feature representation among these methods [2]. A lot of researchers focused on the study of spatial-temporal attention mechanisms, to predict the quality scores for the features of video frames or local regions and obtain aggregated video-level features from both spatial and temporal dimensions. Spatiotemporal Attention Network (STAN) [11] extracted local image features organized by spatial attention model and combined them by temporal attention. Spatial-temporal Clues Integration Module (STIM) [16] mined the spatial-temporal information from features upgraded by refining recurrent unit (RRU). Spatial-Temporal Attention-aware Learning (STAL) [1] focused on the salient parts of persons in videos jointly in both spatial and temporal domains. Limited by the computational memory, these attention mechanisms can only be applied to short clips, then aggregate clip-level features with aver-

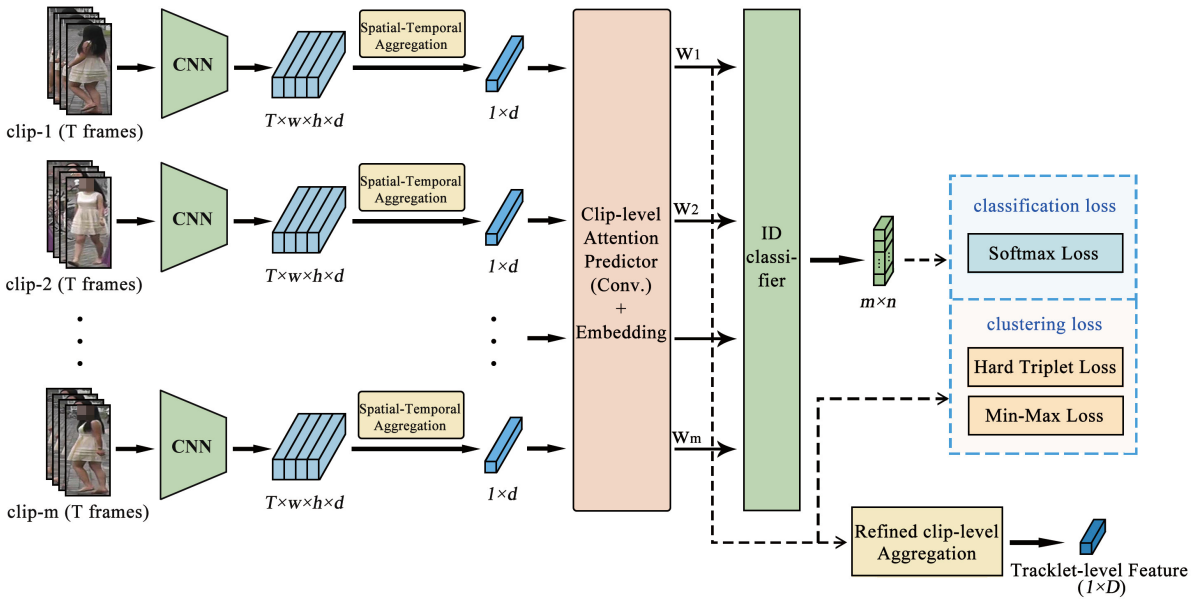

Figure 1. The pipeline of TACAN. Each clip shares the same base model to extract image-level feature maps and then operate aggregation inside the clip. A weight predictor reassigns clip-level attention for each clip so that the model can generate a robust tracklet-level feature. During the training stage, softmax loss is applied as the classification loss, while the combined loss function of triplet loss and min-max loss is applied after the feature embedding layer as the clustering loss.

age/max pooling, which not fully motivated the potential of attention mechanisms.

Other researchers combined adjacent frame features to obtain the integrated feature representation. Region-based Quality Estimation Network (RQEN) [20] learned the partial quality of each image and aggregated the complementary portion of the different frames in an image sequence. Self-and-Collaborative Attention Network (SCAN) [27]generated self and collaborative sequence representations and adopted generalized pairwise similarity measurement to calculate the similarity of video pairs. Moreover, the Spatio-Temporal Completion network (STC-Net) [9] was a GAN-based method. It recovered the appearance of the occluded parts with the Spatio-temporal information, then got better accuracy with the newly generated dataset.

2.3. Metric Learning

In the metric learning task, combining classification loss and verification loss is useful to transfer representations learned from large image classification datasets to fit this identification task [3]. Commonly, the nearest Euclidean feature distance is used when matching, so softmax loss is used for classification and triplet loss is used for metric embedding. Label smoothing strategy [22] was proposed to solve the problem of softmax loss overfitting. Batch-hard triplet loss [6] was more adaptable to mini-batch training and improved the triplet loss, which selected the hardest positive and negative samples in a mini-batch for each anchor sample and computed their metric distances.

Different from existing methods, our work focus on clip-level features aggregation instead of simply aggregation of frames inside clips. We propose a novel TACAN that includes temporal aggregation module and clip attention module. Focusing on higher-level features makes our model express the entire video-level(tracklet-level) features better. Moreover, inspired by Min-max objective [19], we adopt the min-max Loss for video-based Re-ID which can constraint intra-class distances and expand inter-class distances.

3. Method

3.1. Overview

For our proposed TACAN, there are three modules for the entire calculation pipeline: spatial-temporal aggregation within clips, aggregation with clip-level attention and multi-loss function for metric learning. One certain video tracklet S can be divided into several clips $\{C_1, C_2, ..., C_m\}$, each clip has T frames. Inside the clip, on each frame, which is a person bounding box image, we can utilize a CNN model to extract a d-dimensional image-level feature f_{I_i} and the feature set in each clip is $F_I = \{f_{I_1}, f_{I_2}, , ..., f_{I_T}\}$. Then, by aggregating all feature vectors in a certain temporal approach, each clip can be represented by a clip-level feature vector f_{C_i} with the same d dimensions as image-level features. In order to obtain the robust tracklet-level feature vector, the vital module for our proposed method is fusing all

consecutive clip-level features $F_C = \{f_{C_1}, f_{C_2}, , ..., f_{C_m}\}$ with clip-level attention $\mathcal{W} = \{w_1, w_2, ..., w_m\}$. After this refined clip-level aggregation, we can achieve a D-dimensional tracklet-level feature vector f_T. Those feature vectors in all for different video tracklets can be used for person Re-ID by searching the nearest Euclidean distance between two different vectors. According to batch setting during training process, we randomly sample k person ids with j clips for each id to consist a $batch = \{(C_1, ...C_j)_1, (C_1, ...C_j)_2, ..., (C_1, ...C_j)_k\}$. Along with our model, a proposed min-max loss is used, with which we can enlarge the feature vector distance of inter-class and reduce that of intra-class at the same time. Since when aggregating among clips, our model is trained by unsupervised clip-level attention learning, min-max loss function can guide the training process effectively and the model can extract more discriminative tracklet-level features. The pipeline for our proposed method is shown in Figure 1.

3.2. Spatial-Temporal Aggregation within Clips

Considering the limitation of computational memory and probable distraction in overlong sequences, video tracklets are separated down into several clips with a fixed length. Typically, a video tracklet contains tens to hundreds of consecutive image frames, and the aggregation should be operated step by step. Firstly, a CNN model is adopted as an image feature generator. After we get features from all image frames, a rough spatial-temporal aggregation is operated inside the clip.

Existing paper [2] mentioned several temporal aggregation methods after image feature extractor. The generic idea of this rough spatial-temporal aggregation is to make use of video temporal information among several consecutive frames. A well-designed attention mechanism network could perform more outstanding than simple average/max pooling all images in each clip. Figure 2 shows the idea of the temporal attention aggregation method. Attention scores can be obtained from consecutive spatial features by taking temporal information into consideration. For each frame, the d-dimensional spatial feature is f_{I_i} with image attention score a_i, $i \in [1, T]$. So the feature of each clip can be calculated by weighted average $f_C = \frac{1}{T} \sum_{i=1}^{T} a_i \cdot f_{I_i}$. Same as image-level feature vectors, the clip-level feature vectors are also d-dimensional.

3.3. Temporal Aggregation with Clip-level Attention

Ideally, to improve the performance of such spatial-temporal aggregation, we could either utilize a better spatial feature extractor or enrich temporal information with a longer length of clips. Because of the limitation mentioned in Sec. 3.2, the length of clips cannot be set too large. For public video-based Re-ID datasets, such as MARS [13], the frame length of one certain video is not specific, and there is usually no noticeable difference between two consecutive frames. Comparative experiments show that complex temporal modeling brings no remarkable improvement.

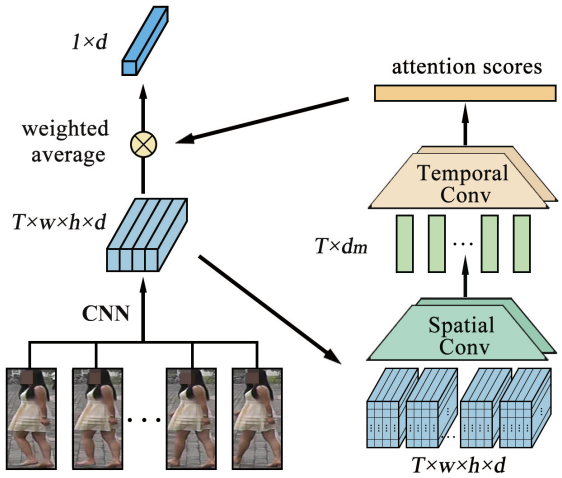

Figure 2. Spatial-Temporal Attention Aggregation Module [2].

To fix this problem, we innovatively propose a refined clip-level aggregation, in which we can obtain more discriminative features by training clip-level attention weights. The clip-level attention can be learned by a weight predictor, then the predicted scores are used for temporal aggregation by weighted average to obtain a discriminative tracklet-level feature. We upgrade the weight-predictor to a higher clip-based level than the image-based level, this allows our model to have more receptive fields and more timing information to score every clip, assigning higher weights to more representative clips. Figure 3 shows how our refined Temporal Aggregation with Clip-level Attention works.

After we obtain all clip-level feature vectors in one video tracklet by the first-step aggregation, we have a set of features $F_C = \{f_{C_1}, f_{C_2}, , ..., f_{C_m}\}$ as input of clip-level aggregation. The second-step aggregation is separated into two branches. The upper branch in Figure 3 is a clip-level weight predictor as mentioned before. In detail, this layer is a FC Layer. Each input feature vector produces a corresponding clip-level weight through this layer:

$$\begin{aligned} \mathcal{W} &= W \cdot F_C \\ &= W \cdot \{f_{C_1}, f_{C_2}, , ..., f_{C_m}\} \\ &= \{W \cdot f_{C_1}, W \cdot f_{C_2}, , ..., W \cdot f_{C_m}\} \\ &= \{w_1, w_2, ..., w_m\} \end{aligned} \tag{1}$$

In our clip-level aggregation, d-dimensional clip-level feature vectors from first-step rough aggregation are regarded as middle-stage features and directly used for clip-level attention generation.

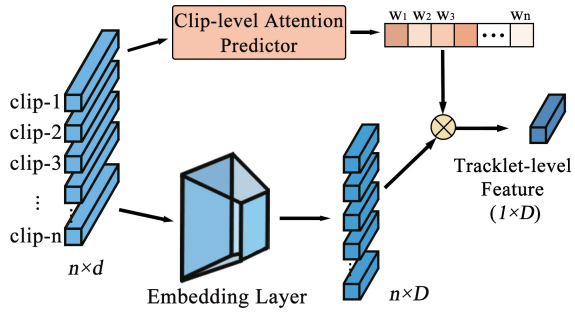

Figure 3. Refined Aggregation with Clip-level Attention

The lower branch is the embedding layer. In order to generate the final features, an embedding layer is added to increase the network depth and to obtain higher-level features. Meanwhile, it can decrease the dimensions of the final tracklet-level feature vectors. We denote the embedding transform as $\mathcal{T}(\cdot)$ with D-dimensional output and the whole Embedding Layer can be expressed as:

$$\begin{aligned} E &= \mathcal{T}(F_C) \\ &= \mathcal{T}(\{f_{C_1}, f_{C_2}, , ..., f_{C_m}\}) \\ &= \{\mathcal{T}(f_{C_1}), \mathcal{T}(f_{C_2}), ..., \mathcal{T}(f_{C_m})\} \\ &= \{f_{\mathcal{E}_1}, f_{\mathcal{E}_2}, , ..., f_{\mathcal{E}_m}\} \end{aligned} \tag{2}$$

With output from two branches, we operate the second-step refined temporal aggregation. The temporal aggregation with clip-level attention $\mathcal{W}$ can be operated by the weighted average:

$$\begin{aligned} \boldsymbol{f} &= \frac{1}{m}(\mathcal{W} \cdot E) = \frac{1}{m}\sum_{i=1}^{m} w_i f_{\mathcal{E}_i} \\ &= \frac{1}{m}(w_1 f_{\mathcal{E}_1} + w_2 f_{\mathcal{E}_2} + ... + w_m f_{\mathcal{E}_m}) \end{aligned} \tag{3}$$

where $\mathcal{W}$ is the set of clip-level attention weights. E is the embedded features set. m is the number of clips divided from one tracklet and $\boldsymbol{f}$ is the final D-dimensional tracklet-level feature vector.

3.4. Multi-loss Function

As is described in Sec. 3.3, clip-level attention is achieved by an unsupervised strategy because there is no annotation of attention for video clips. Clip-level attention is automatically learned by our model. To make this strategy work out during the training process, a novel multi-loss function is designed. Figure 1 shows how the pipeline works during the training process, from which we can notice that the unit for training is a video clip. Guided by our loss function, the clip-level attention can be optimized to make training samples separated among different classes and clustered within the same classes.

The overall loss function of our proposed TACAN model can be formulated as:

$$\mathcal{L} = \mathcal{L}_s + \mathcal{L}_t + \mathcal{L}_m \tag{4}$$

Softmax loss $\mathcal{L}_s$ is used as classification loss. With the supervision of person IDs, softmax loss guides the network to generate accurate classification. Another part of our loss function is to make constraints on feature distance. It is a combination of hard triplet loss $\mathcal{L}_t$ and min-max objective $\mathcal{L}_m$. Triplet loss can be formulated as:

$$\mathcal{L}_t = \sum_{i=1}^{N}\left[\alpha + \|\boldsymbol{f_i^a} - \boldsymbol{f_i^p}\|_2^2 - \|\boldsymbol{f_i^a} - \boldsymbol{f_i^n}\|_2^2\right]_+ \tag{5}$$

where $\boldsymbol{f_i^a}$, $\boldsymbol{f_i^p}$, $\boldsymbol{f_i^n}$ are the anchor feature, positive feature and negative feature, respectively. To improve performance after introducing triplet loss, we use the batch hard positive and negative features, which means:

$$\begin{cases} \boldsymbol{f_i^p} = \arg\max_{f_i^p} \|\boldsymbol{f_i^a} - \boldsymbol{f_i^p}\|_2^2 \\ \boldsymbol{f_i^n} = \arg\min_{f_i^n} \|\boldsymbol{f_i^a} - \boldsymbol{f_i^n}\|_2^2 \end{cases} \tag{6}$$

Inspired by the work [19], we propose the min-max loss as an extension in this part. Min-max loss aims to directly minimize the intra-class distances and maximize the inter-class distances among the feature vectors after the embedding layer. We use $\boldsymbol{m_k}$ to denote the mean vector of class k with same person ID and use $\boldsymbol{m}$ to denote the overall mean vector of a batch. The calculation of the intra-class distance $\boldsymbol{S_k^{(I)}}$ for class k can be denoted as:

$$\boldsymbol{S_k^{(I)}} = \sum_{i \in \pi_k} (\boldsymbol{f_i} - \boldsymbol{m_k})^\top (\boldsymbol{f_i} - \boldsymbol{m_k}) \tag{7}$$

The total intra-class distance $\boldsymbol{S^{(I)}}$ and total inter-class distance $\boldsymbol{S^{(B)}}$ can be formulated as:

$$\begin{cases} \boldsymbol{S^{(I)}} = \sum\limits_{k=1}^{K} \boldsymbol{S_k^{(I)}} \\ \boldsymbol{S^{(B)}} = \sum\limits_{k=1}^{K} n_k (\boldsymbol{m_k} - \boldsymbol{m})^\top (\boldsymbol{m_k} - \boldsymbol{m}) \end{cases} \tag{8}$$

Min-max loss aims at minimizing the intra-class distance $\boldsymbol{S^{(I)}}$ while maximizing the inter-class distance $\boldsymbol{S^{(B)}}$. So the loss function can be defined as:

$$\mathcal{L}_m = \frac{\boldsymbol{S^{(I)}}}{\boldsymbol{S^{(B)}}} \tag{9}$$

Overall, $\mathcal{L}_s$ is the softmax loss for classification. $\mathcal{L}_t$ is the triplet loss which can enlarge feature distances among different classes while $\mathcal{L}_m$ is the min-max objective function which can compress features within identical classes. Since during the testing stage we search for features with the nearest feature distance as matches, the combination of $\mathcal{L}_t$ and $\mathcal{L}_m$ can definitely enhance the model performance.

Method	mAP	Rank-1	Rank-5	Rank-20
baseline(ResNet50+$\mathcal{L}_s$)	61.4	72.6	87.2	93.2
ResNet50+$\mathcal{L}_s$+$\mathcal{L}_t$	76.2	82.6	93.9	96.8
ResNet50+TA+$\mathcal{L}_s$+$\mathcal{L}_t$	76.8	83.9	93.9	97.2
ResNet50+TA+CA+$\mathcal{L}_s$+$\mathcal{L}_t$	81.6	86.3	95.4	98.1
ResNet50+TA+CA+$\mathcal{L}$	83.0	88.2	96.0	**98.3**
SENet50+TA+CA+$\mathcal{L}$	**84.0**	**89.1**	**96.1**	98.0

Table 1. Module-wise performance on the MARS dataset. ST is a shorthand for spatial-temporal aggregation module, and CA is a shorthand for clip-level attention module.

4. Experiments

This section reports the experimental results of our proposed TACAN model. Experiments are mainly conducted on the MARS [13] dataset. Firstly, we introduce the MARS dataset. Then through comparative experiments, we show how each component in the proposed TACAN model improves the overall performance. Next we also report results based on three additional datasets, the iLIDS-VID dataset [24], PRID-2011 dataset [7] and DukeMTMC-VideoReID [18], to validate the robustness of our method. Finally, we mention the implementation details.

4.1. Datasets and Measurement

MARS contains 20,478 traclets of 1261 person IDs which are captured at Tsinghua University campus from 6 non-overlapping camera views. The dataset is divided into a training set with 625 person IDs and a testing set with 626 person IDs. There are 8,298 tracklets for the training set and 12,180 tracklets (1,980 in query and 9,330 in gallery) for the testing set. As an extension of the Market-1501 dataset [28], MARS is one of the largest public video-based person re-identification datasets.

In the following experiments, the model accuracy is measured by the Cumulative Matching Characteristic (CMC) table and the mean average precision (mAP) score. We evaluate all models on the MARS dataset to keep these measurements the same as that from the original MARS dataset released [13].

4.2. Performance Comparison of Different Modules

The baseline approach contains only ResNet50 model trained by softmax loss $\mathcal{L}_s$. Both image-level feature aggregation and clip-level feature fusion are based on average pooling. Table 1 shows the module-wise performances on the MARS dataset.

As can be seen from Table 1, after applying $\mathcal{L}_s$ and $\mathcal{L}_t$, the mAP score and the Rank-1 score were improved by 14.8% and 10.0% respectively. When we added the TA module, the model had a slight performance boost. When the CA module was added to the network, the representation capacity was significantly improved where the mAP was increased to 81.6% and the Rank-1 was increased to 86.3%. After applying the multi-loss $\mathcal{L} = \mathcal{L}_s + \mathcal{L}_t + \mathcal{L}_m$, we achieved 83.0% mAP score and 88.2% Rank-1 score. Finally, when we adopted the SENet50 model as our backbone to complete our proposed TACAN framework, we obtained an additional 1% performance improvement on both mAP and Rank-1 scores, and the performance outperformed all listed benchmark methods as shown in Table 2 below. It is seen from Table 2 that, our network outperforms the best literature [9] by 0.6% in Rank-1 score, and 1.7% by the mAP score.

Method	mAP	Rank-1	Rank-5	Rank-20
ASTPN[26]	-	44	70	81
JSTRN[31]	50.7	70.6	90.0	97.6
STAN[11]	65.8	82.3	-	-
RQEN[20]	71.1	77.8	88.8	94.3
STIM[16]	72.7	84.4	93.2	96.3
SCAN[27]	76.7	86.6	94.8	97.1
TA[2]	76.7	83.3	93.8	97.4
ResNet3D NL[14]	77.0	84.3	94.6	-
STAL[1]	73.5	82.2	92.8	98.0
VRSTC[9]	82.3	88.5	**96.5**	-
TACAN(ResNet)	83.0	88.2	96.0	**98.3**
TACAN(SEnet)	**84.0**	**89.1**	96.1	98.0

Table 2. The comparison with existing video-based Re-ID approaches on MARS. TACAN using ResNet50 as backbone can already reach the state-of-the-art, but using SEnet50 will achieve better performance. Therefore, we recommend using SEnet as the backbone.

4.3. The Visualization of Model Effects

Since the proposed method in this paper uses an unsupervised strategy to learn clip-level attention. This clip-level attention is automatically learned by our model and it denotes the weight of one certain video clip that can be used for tracklet feature aggregation. We expect the model can give out a higher clip-level attention value for clips with better quality.

In our experiment setting, the sequence length for one clip was 4, which means there were 4 images in a video clip. In Figure 4, each sub-figure contains three consecutive clips picked from a certain tracklet of MARS dataset. From this figure, we clearly see that our model can give a higher score when the pedestrian bounding boxes are visible without any occlusion. In the opposite case, the score is decreased to weaken the effects of features with occlusion. In these two examples, main bodies for re-identification are overlapped with other person. When the overlapping happens, the model gives lower clip-attention values.

In order to show the effectiveness of clip-level attention(CA) more clearly, Figure 6 takes a longer tracklet as an example to illustrate how CA works on a tracklet. Comparing clip2, clip15, clip20, clip30, clip34, we can see that

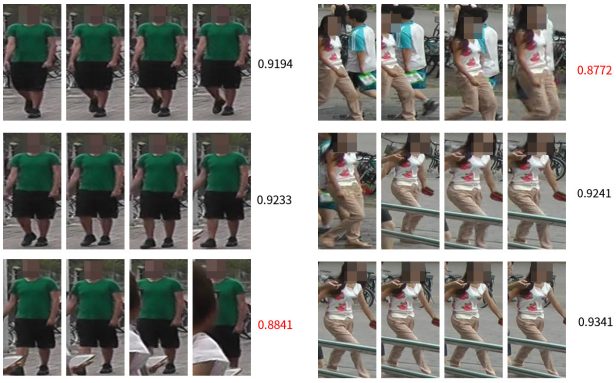

Figure 4. Consecutive video clips with corresponding clip-level attention values given out by pre-trained model.

CA is not sensitive to background changes, which means the focus of CA is on the foreground(pedestrian) rather than the background. Comparing clip21 to clip30, we can see that when severe occlusion occurs, clip-level attention is descending to weaken the effect of the negative sample on the final feature vector. When the occlusion disappears, the CA is increased to allow subsequent positive sample features to have more effect. It shows that by introducing clip-level attention, our model can give out more representative feature vectors since the final tracklet-level features are more correlated with good quality video clips which can obtain higher clip-level attention for our model.

To visualize how min-max loss influenced the model performance, 5 person IDs (121 tracklets in total) were randomly selected from MARS. We used pre-trained models to extract features from these tracklets. With t-SNE [17] algorithm, we could obtain 2-dimensional features which made it easier to visualize. By introducing min-max loss in our proposed method, it was expected that features can be more clustered in feature space. Figure 5 shows the visualization of these tracklet-level features represented by two different models, which were trained without/with min-max loss.

Table 3 shows the quantification of the clustering effect. We used the Largest Euclidean Distance in a cluster to measure the clustering effect:

$$diam(C) = \max_{1<=i<j<=|C|}(dist(\boldsymbol{x_i}, \boldsymbol{x_j})) \qquad (10)$$

where $|C|$ is the number of samples in one cluster, $dist(\cdot)$ here is Euclidean distance and $(\boldsymbol{x_i}, \boldsymbol{x_j})$ are two different features vectors in the feature space. The features are more clustered with a smaller Largest Euclidean Distance. It is clear that extracted by the model with min-max loss, features are gathered much more together and more discriminative for the same identity.

4.4. Extended Datasets Validation

PRID-2011[7] consists of 400 video sequences of 200 people captured at uncrowded outdoor scenes by 2 non-overlapping camera views. The video sequences are between 5 and 675 frames in length and have an average of 100. Because of the simple and clean background and rare cluttered occlusions, PRID-2011 is relatively less challenging than the MARS dataset.

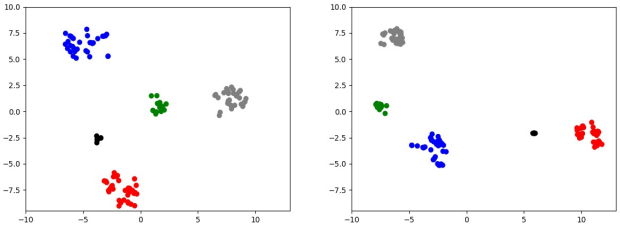

Figure 5. Clustering effect visualization. Using the t-SNE algorithm to reduce feature dimensions. The left image used the model trained without min-max loss. The right image used the model trained with min-max loss.

samples	without min-max loss	with min-max loss
sample1	0.9204	0.8882
sample2	0.7038	0.5321
sample3	0.8364	0.6553
sample4	0.8448	0.6056
sample5	0.8931	0.6629

Table 3. Clustering effect quantification. Using the Largest Euclidean Distance in a cluster to measure the clustering effect.

iLIDS-VID[24] contains 600 video sequences of 300 people captured at an airport arrival lobby from 2 non-overlapping camera views. The video sequences are between 23 and 192 frames in length and have an average of 73. This dataset is more challenging due to clothing similarities among people, lighting and viewpoint variations across camera views, cluttered background and occlusions.

Both of the datasets are randomly half-half separated into training and testing sets. Because these two datasets were captured by only two cameras, during testing, data from one camera is regarded as query and the other as gallery. We repeated the separation of training and testing process ten times and averaged the results into the final score as shown in Table 4.

DukeMTMC-VideoReID[25] is a subset of the DukeMTMC tracking dataset [18]. Training set consists of 702 person IDs with 2196 tracklets. Testing set is also separated into query and gallery. There are 702 person IDs with one tracklet for each ID in query set and there are 1110 person IDs with 2636 tracklets in total for gallery set.

It can be seen from Table 4 and Table 5 that, our method outperforms the best existing method by 3.1% on mAP

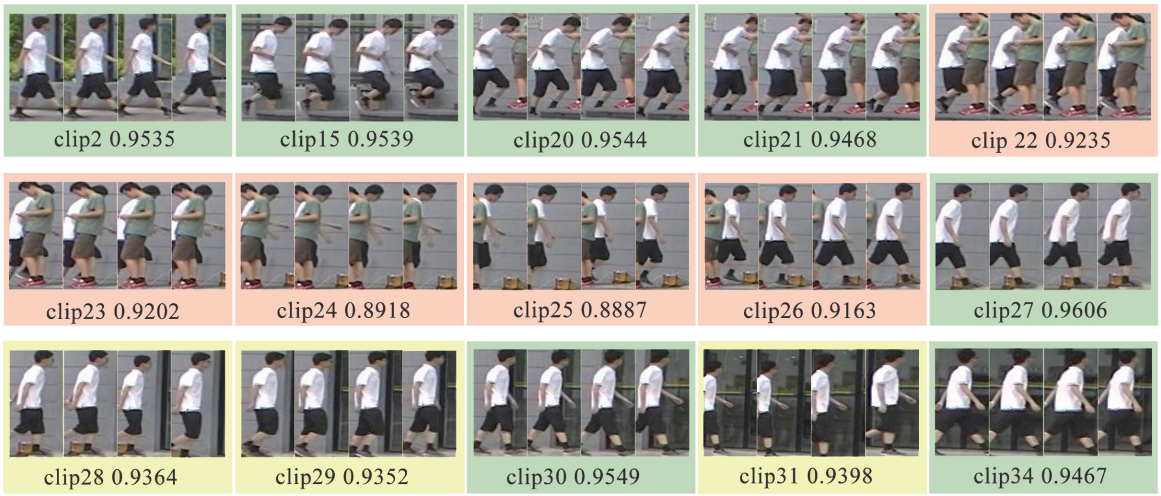

Figure 6. The visualization of one certain tracklet from MARS. Only 15 representative clips are selected and showed instead of the whole tracklet(190 frames in it). We use different colors to mark different situations. Green denotes the CA scores are more than 0.9460 and has a high level of effect for the final feature. Yellow denotes the CA scores are between 0.9240 and 0.9460, which have the medium level effect. Red denotes the CA scores are less than 0.9240, which means occlusion occurs on corresponding clips and such clips have less effect on the final feature vectors.

Method	iLIDS-VID		PRID-2011	
	mAP	Rank-1	mAP	Rank-1
ASTPN[26]	-	62	-	77
JSTRN[31]	-	55.2	-	79.4
STAN[11]	-	80.2	-	93.2
RQEN[20]	-	80.0	-	93.4
STIM[16]	-	84.3	-	92.7
SCAN[27]	89.9	88.0	95.8	95.3
ResNet3D NL[14]	-	81.3	-	91.2
STAL[1]	-	82.8	-	92.7
VRSTC[9]	-	83.4	-	-
TACAN(ours)	**93.0**	**88.9**	**96.7**	**95.3**

Table 4. The comparison with existing video-based Re-ID approaches on iLIDS-VID and PRID-2011

Method	mAP	Rank-1	Rank-5	Rank-10
EUG[25]	78.3	83.6	94.6	97.6
VRSTC[9]	93.5	95.0	99.1	99.4
TACAN(ours)	**95.4**	**96.2**	**99.4**	**99.6**

Table 5. The comparison of metrics on DukeMTMC-VideoReID

score and 0.8% on Rank-1 score for iLIDS-VID, and by 0.9% on mAP score for PRID-2011. For DukeMTMC-VideoReID, our method is also better than other methods on mAP and all CMC scores. These results validate the effectiveness and robustness of our proposed approach.

4.5. Implementation Details

We use the image feature extractor models pre-trained on the ImageNet dataset. Both ResNet50 [5] and SENet50 [10] are used in our experiments. Each input video frame is resized to 224×112 pixels. All training and testing are performed on a single GPU. Limited to computing resources, the clip length T is set to 4 while the batch size n is set to 32. The number of instances with the same identity is set to 4 which means we randomly sample 4 sets of tracklets each pedestrian for the training stage. The dimension of middle-stage feature (clip-level feature) $d = 2048$, and the dimension of embedding tracklet-level feature is $D = 1024$. The network is updated by stochastic gradient descent algorithm. The initial learning rate is 0.0003 with learning rate decay strategy during training.

5. Conclusions

In this paper, we proposed a novel Temporal Aggregation with Clip-level Attention Network (TACAN) which is an end-to-end CNN model for video-based person Re-ID. TACAN shows better accuracy benefiting from temporal aggregation and clip-level attention, where clip-level attention is learned automatically from each clip of frames and then used to weighted combine the clip-level feature into a final tracklet-level representation. In addition, we adopt the min-max loss for video-based Re-ID along with hard triplet loss, which makes the training process more effective. Experiments show that our proposed TACAN model reaches superior performance on four popular benchmarks over existing state-of-the-art methods.

References

[1] G. Chen, J. Lu, and J. Zhou. Spatial-temporal attention-aware learning for video-based person re-identification. *IEEE Transactions on Image Processing*, pages 4192–4205, 2019.

[2] J. Gao and R. Nevatia. Revisiting temporal modeling for video-based person reid. *arXiv preprint arXiv:1805.02104*, 2018.

[3] M. Geng, Y. Wang, X. Tao, and Y. Tian. Deep transfer learning for person re-identification. *ArXiv e-prints*, 2016.

[4] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, pages 2672–2680, 2014.

[5] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[6] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. *ArXiv e-prints*, 2017.

[7] M. Hirzer, C. Beleznai, P. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In *Image Analysis*, pages 91–102, 2011.

[8] E. Hoffer and N. Ailon. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*, pages 84–92. Springer, 2015.

[9] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen. Vrstc: Occlusion-free video person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7183–7192, 2019.

[10] H. Jie, S. Li, S. Albanie, S. Gang, and E. Wu. Squeeze-and-excitation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[11] S. Li, S. Bak, P. Carr, and X. Wang. Diversity regularized spatiotemporal attention for video-based person re-identification. In *CVPR*, 2018.

[12] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 152–159, 2014.

[13] Z. Liang, B. Zhi, Y. Sun, J. Wang, S. Chi, S. Wang, and T. Qi. Mars: A video benchmark for large-scale person re-identification. In *Computer Vision – ECCV 2016*, 2016.

[14] X. Liao, L. He, Z. Yang, and C. Zhang. Video-based person re-identification via 3d convolutional networks and non-local attention. In *Computer Vision – ACCV 2018*, 2019.

[15] Y. Liu, J. Yan, and W. Ouyang. Quality aware network for set to set recognition. In *CVPR*, pages 4694–4703, 2017.

[16] Y. Liu, Z. Yuan, W. Zhou, and H. Li. Spatial and temporal mutual promotion for video-based person re-identification. In *AAAI*, 2019.

[17] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[18] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision workshop on Benchmarking Multi-Target Tracking*, 2016.

[19] W. Shi, Y. Gong, X. Tao, J. Wang, and N. Zheng. Improving cnn performance accuracies with min–max objective. *IEEE transactions on neural networks and learning systems*, 29(7):2872–2885, 2017.

[20] G. Song, B. Leng, Y. Liu, C. Hetang, and S. Cai. Region-based quality estimation network for large-scale person re-identification. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[21] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 480–496, 2018.

[22] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016.

[23] G. Wang, Y. Yuan, C. Xiong, J. Li, and Z. Xi. Learning discriminative features with multiple granularities for person re-identification. *ArXiv e-prints*, 2018.

[24] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by video ranking. In *Computer Vision – ECCV 2014*, 2014.

[25] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Ouyang, and Y. Yang. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[26] S. Xu, C. Yu, G. Kang, Y. Yang, S. Chang, and Z. Pan. Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In *IEEE International Conference on Computer Vision*, 2017.

[27] R. Zhang, H. Sun, J. Li, Y. Ge, L. Lin, P. Luo, and X. Wang. Scan: Self-and-collaborative attention network for video person re-identification. *IEEE Trans. on Image Processing*, 2019.

[28] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *Computer Vision, IEEE International Conference on*, 2015.

[29] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3754–3762, 2017.

[30] S. Zhou, J. Wang, D. Meng, Y. Liang, Y. Gong, and N. Zheng. Discriminative feature learning with foreground attention for person re-identification. *IEEE Transactions on Image Processing*, 2019.

[31] Y. Zhou, Zhen adn Huang, W. Wang, L. Wang, and T. Tan. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In *CVPR*, 2017.