Clip-level Feature Aggregation: A Key Factor for Video-based Person Re-Identification

Chengjin Lyu(✉), Patrick Heyer-Wollenberg, Ljiljana Platisa, Bart Goossens, Peter Veelaert, and Wilfried Philips

TELIN-IPI, Ghent University - imec, 9000 Ghent, Belgium
chengjin.lyu@ugent.be

Abstract. In the task of video-based person re-identification, features of persons in the query and gallery sets are compared to search the best match. Generally, most existing methods aggregate the frame-level features together using a temporal method to generate the clip-level features, instead of the sequence-level representations. In this paper, we propose a new method that aggregates the clip-level features to obtain the sequence-level representations of persons, which consists of two parts, i.e., Average Aggregation Strategy (AAS) and Raw Feature Utilization (RFU). AAS makes use of all frames in a video sequence to generate a better representation of a person, while RFU investigates how batch normalization operation influences feature representations in person re-identification. The experimental results demonstrate that our method can boost the performance of existing models for better accuracy. In particular, we achieve 87.7% rank-1 and 82.3% mAP on MARS dataset without any post-processing procedure, which outperforms the existing state-of-the-art.

Keywords: Person re-identification · Convolutional neural network · Feature aggregation

1 Introduction

Person re-identification refers to identifying a person of interest across different images/video sequences, by comparing the tracks of persons from multiple cameras with non-overlapping areas. Recently, it has drawn increasing attention thanks to its potential applications in intelligent video surveillance, such as person tracking [7] and search [28]. It is still a very challenging task due to occlusion, background clutter, as well as the intensive changes in lighting, pose and viewpoint.

Technically speaking, person re-identification tasks could be divided into two sub-tasks: image-based and video-based re-identification. Single image based re-identification methods have achieved impressive results. However, image-based results may be seriously affected by the quality of images, especially when there is significant occlusion which may even lead to false detection on a single frame. Different from image-based methods, video-based re-identification approaches

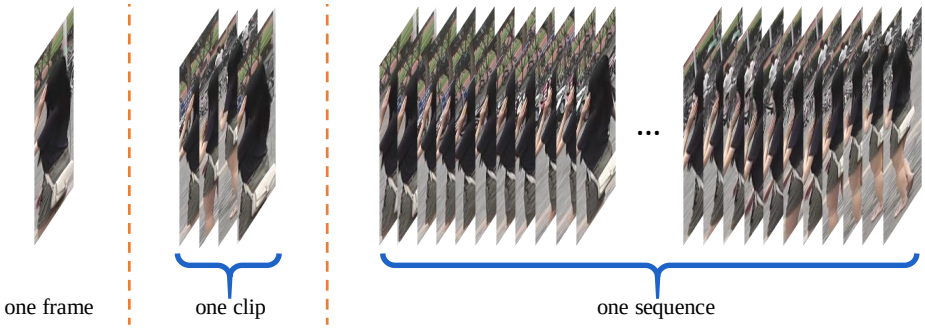

Fig. 1. Examples of a frame, a clip and a sequence for the same person. A clip is defined as a few frames of fixed length. A sequence contains all the available frames in a track.

use multiple frames from a complete video sequence to extract the features of a person. It is natural and practical that a sequence of images can provide richer information compared to a single image [33]. The appearances of a person from multiple frames are complementary to each other, which could be used to extract more robust features. Moreover, if a person is moving in a given sequence, spatiotemporal information that describes a person's motion can be summarized from a sequence.

The main idea of video-based person re-identification approaches is to design a spatial-temporal model to extract discriminative representations for video sequences. In general, a typical video sequence lasts a few or even dozens of seconds in person re-identification, which leads to dozens or even hundreds of frames in a sequence. However, due to the physical limitation of GPU's memory, it is almost not realistic to feed all the frames of a sequence into the spatial-temporal model and train the model in a batch with enough sequences. Furthermore, considering the trade-off between the training cost and the information contained within each subsequence (named as clip in this paper), clips of fixed length (usually 4, 6, or a little more) are usually utilized for further feature extraction, which also makes the model suitable for sequences of variable lengths [29]. Examples of a frame, a clip and a sequence for the same person are shown in Fig. 1. Thus, most existing video-based person re-identification approaches focus on the designs of spatial-temporal models for clip-level feature extraction. However, the effect of clip-level feature aggregation has not been studied well.

To explore this issue, we propose a simple yet powerful method to aggregate clip-level features. It is parameter-free, and could be used to boost the performance of existing video-based re-identification methods. Our main contributions in this work are as follows:

1. Propose an Average Aggregation Strategy (AAS) to aggregate clip-level features, where clip-level feature vectors from the same sequence are averaged to represent the whole sequence and then cosine similarities of the averaged vectors are computed for the inference stage.

2. Investigate the influence of the final batch normalization layer for video-based person re-identification, and demonstrate Raw Feature Utilization (RFU) for both training and inference stages.
3. Perform experimental evaluation on MARS dataset. The results show that our proposed method outperforms the existing state-of-the-art without any post-processing procedure.

2 Related Work

In this section, we review related researches including image-based person re-identification, video-based person re-identification, and feature aggregation used in person re-identification.

Image-based person re-identification was first proposed by Gheissari et al. [10] in 2006. In that paper, a combination of color and salient edgel histograms was utilized to perform visual matching on a 44-person dataset. According to the methods of feature generation, the research progress of image-based person re-identification can generally be divided into hand-craft and deep-learning based methods. Hand-craft features such as texture histograms [11], SIFT descriptor [31], and color histograms [5] are widely used. In recent years, with the great success of convolutional neural network (CNN), deep-learning based methods have made distinguished achievements on image-based person re-identification tasks. A siamese model was first chosen to train a CNN for person re-identification with pairs of images [16]. Cheng et al. [3] use triplets of images to train a CNN which can learn both global and local features. Varior et al. [24] introduced long short-term memory (LSTM) module into siamese network to extract spatial relationships. Sun et al. [23] proposed a part-based convolutional baseline (PCB) network which extracts a feature vector consisting of several part-level features. Luo et al. [19] collected and evaluated some effective training tricks in image-based person re-identification.

Video-based person re-identification appeared in the year of 2010 [1, 8], where it was first named as multi-shot person re-identification. At first, best match between sets of image-level features was explored. Karaman et at. [14] introduced a conditional random field (CRF) to person re-identification via building a neighborhood topology based on spatial and temporal similarity. Cho et al. [4] proposed an approach to estimate target poses and perform multi-pose model generation and matching. In the second stage, hand-craft features were designed for spatial-temporal information extraction. In [26], spatiotemporal features were first introduced to video-based person re-identification. In [18], body-action units are extracted and then fed to Fisher vectors for final feature generation. More recently, thanks to the emergence of large-scale datasets (e.g., iLIDS-VID [27] and MARS [32]), deep-learned methods have improved the performance of video-based re-identification rapidly. McLaughlin et al. [20] employed a recurrent neural network (RNN) to summarize features from CNNs. Liu et al. [17] built a network which can accumulate motion context from adjacent frames by with the help of RNN. To better extract spatial-temporal information

from the videos, attention mechanisms are introduced to this community. Zhou et al. [35] trained a network to pick out the most discriminative frames with the help of a temporal attention model. Song et al. [21] employed a landmark detector and fully convolutional network to generate region-based quality and representation. Li et al. [15] proposed a multiple attention framework to discover a diverse set of distinctive body parts. Fu et al. [9] introduced a spatial-temporal attention (STA) approach to generate robust clip-level feature representation for video-based person re-identification. Su et al. [22] proposed a k-reciprocal harmonious attention network (KHAN), where spatial and channel attentions are fused as spatial attention and k-reciprocal attention is calculated for temporal attention.

Feature aggregation in person re-identification. In the first works of video-based person re-identification [1, 8], multi-match strategies were utilized which take a sequence of images as multiple features for matching. Generally, these multi-match strategies may lead to high computational cost and poor scalability. Therefore, current video-based person re-identification methods usually employ an aggregation step to generate one single feature vector for a video sequence. This step could be max or average pooling [20, 17], learned by LSTM [29] or reinforcement learning (RL) [30], for frame-level features. These methods have reached impressive results. What is more, for a long video sequence, the clip-level feature extraction methods could output multiple features to represent the same person. The recent work of [2] introduced a way of competitive snippet-similarity aggregation, which is a variant of multi-match strategy based on clip-level features. However, the aggregation of these clip-level features in video-based person re-identification has not been studied well. In this paper, we propose a roust method for clip-level feature aggregation and assess its performance with the aim to ensure efficient use of the information contained in a video sequence.

3 Methodology

Given an existing model which can be used to extract clip-level spatial-temporal features, where and how to aggregate these features are both essential questions for feature aggregation. In this section, a simple but effective aggregation method which boosts the performance of clip-level features for video-based person re-identification is presented.

3.1 Baseline Model

In this paper, we adopt the STA (Spatial-Temporal Attention) [9] model as our baseline model (see Fig. 2). ResNet50 [12] is chosen as backbone network, which is also the backbone network for many video-based re-identification works. It is worth noting that our method is not restricted to the STA model, and could also work with other video-based based person re-identification methods with similar structure. The calculation progress of spatial-temporal attention in the STA model is listed as follows.

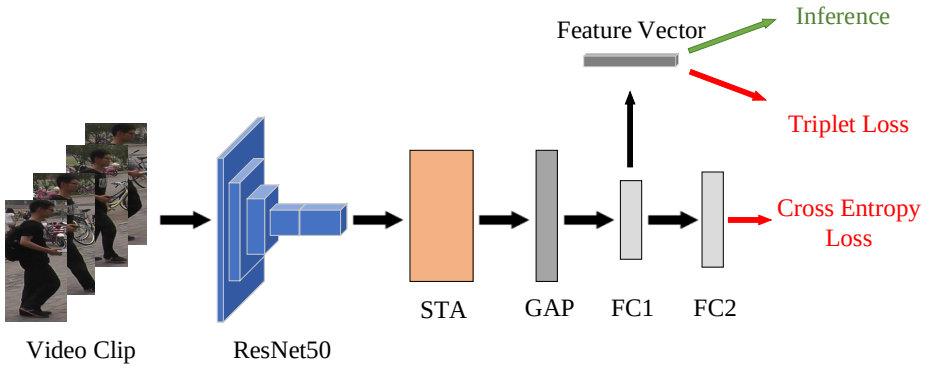

Fig. 2. Pipeline of the baseline model. An input video clip is generated via random sampling. The output is a feature vector that represents the input person.

Suppose that the input clip of length N is represented as $\{I_n\}_{\{n=1:N\}}$, where I_n is a whole-body image of a person. The backbone network generates a set of $H \times W$ feature maps $\{f_n\}_{\{n=1:N\}}$, with $D = 2048$ channels. The attention map a_n is calculated as:

$$a_n(h,w) = \frac{\| \sum_{d=1}^{D} f_n(h,w,d)^2 \|_2}{\sum_{h=1}^{H} \sum_{w=1}^{W} \| \sum_{d=1}^{D} f_n(h,w,d)^2 \|_2}, \tag{1}$$

where H, W are the height and width of the feature maps, respectively. Then, motivated by the success of part-level features [23], divide these feature and attention maps into M blocks horizontally:

$$\begin{cases} f_n = [f_{n,1}, ..., f_{n,m}, ..., f_{n,M}] \\ a_n = [a_{n,1}, ..., a_{n,m}, ..., a_{n,M}]. \end{cases} \tag{2}$$

After that, the attention score for the mth horizontal region on the nth frame is computed as:

$$s_{n,m} = \sum_{i,j} \| a_{n,m}(i,j) \|_1, \tag{3}$$

where $\{(i,j)\}$ covers the mth horizontal region on the nth frame. Then, an $N \times M$ attention matrix $S = [s_{n,m}]_{N \times M}$ is obtained, which stores the attention scores for different horizontal regions on different frames. The final spatial-temporal attention scores can be calculated as:

$$S(n,m) = \frac{s_{n,m}}{\sum_n \| s_{n,m} \|_1}. \tag{4}$$

Finally, element-wise multiplication of the attention score $S(n,m)$ and horizontal region $f_{n,m}$ is employed to generate the final feature maps.

To obtain the feature vector that represents the input video clip, a global average pooling (GAP) followed by a fully connected (FC) layer is used. For the training progress, the combination of batch-hard triplet loss [13] and cross entropy loss is adopted.

Fig. 3. Examples of random sampled images from a sequence. (a) Ideal case that all sampled images are of good quality, (b) Worst case that all sampled images are badly occluded.

3.2 Average Aggregation Strategy

In the baseline model, every video sequence is reduced to N frames via random sampling for both the training and inference stages. Thus, there is no feature aggregation operation on the clip-level features at the stage of inference, for there is only one clip for a sequence. However, representation for a whole sequence using only one N-frame clip is somewhat risky. Sometimes, a clip generated by random sampling could suffer from the low-quality samples, which might lead to false match at the stage of inference. Examples of random sampled images can be found in Fig. 3.

To reduce the uncertainty of a single clip, we use multiple random sampled clips to generate the final feature representation for the whole sequence. The sampled clips from the same sequence are not overlapped with each other. Let v_i denote an arbitrary output feature clip-level vector of the baseline model, the average of all clip-level feature vectors in the same sequence is:

$$\bar{v} = \frac{1}{C} \sum_{i=1}^{C} v_i, \tag{5}$$

where C is the amount of sampled clips, and $\bar{v}$ is the obtained sequence-level feature vector.

Motivated by the successful work of face verification [25], at the stage of inference, the cosine similarity is used to calculated the similarity of two feature vectors, so as to find the best match. Different from Euclidean distance, the cosine similarity only counts on the angle between two vectors instead of their magnitudes in the high-dimensional feature space. The influence of AAS is discussed in Section 4.

3.3 Raw Feature Utilization

A batch normalization (BN) layer is widely used in various deep neural networks to speed up the training progress and also improve the performance of networks. In our re-implementation of STA model, we also include a BN layer (see Fig. 4).

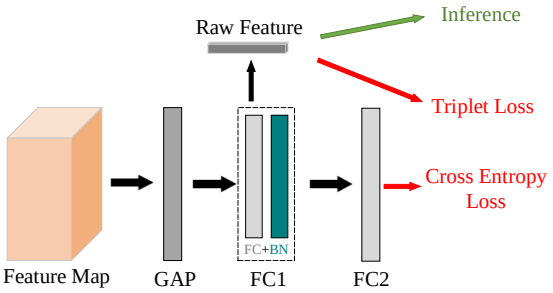

Fig. 4. Illustration of the proposed raw feature utilization. The feature vector before the batch normalization layer is defined as raw feature. This raw feature is utilized for further feature aggregation and the computation of triplet loss for training.

However, in the task of video-based person re-identification, the output of a feature extraction model are feature vectors and then perform future matching based on these feature vectors, which means the direct output of the deep models are feature vectors instead of labels of persons. This is slightly different from the classic computer vision tasks (e.g., image classification and segmentation), where the labels are exactly the outputs of deep neural networks.

Nevertheless, for a person re-identification task, the batch normalization operation of a final feature vector might have harmful effects, since the original distribution of data get changed after batch normalization [19]. In this paper, this change of distribution could influence the performance of subsequent feature aggregation for inference and the computation of the triplet loss function for training. In order to keep the advantage of the final batch normalization layer and also reduce its side effects, we define a way of utilization of raw features in video-based person re-identification for both training and inference stages.

Let v_i^r denote an arbitrary feature vector which is a clip-level representation of a person. It comes after the first fully connect (FC1 in Fig. 4) layer directly, without the final BN operation, which means that v_i^r keeps the original distribution in the feature space. Thus, we name v_i^r as a raw feature. In this paper, we explore the utilization of raw features in both inference and training stages (as shown in Fig. 4). For the inference stage, the final sequence-level feature vector is calculated based on $\{v_i^r\}$ using our average aggregation strategy in Section 3.2. For the training stage, raw features are used to compute the triplet loss. The experimental analysis of raw feature utilization is demonstrated in Section 4.

4 Experiments

4.1 Experimental Setup

Dataset. MARS dataset [32] is the largest video-based person re-identification dataset, which contains 20,715 video sequences. Every sequence has more than 59 frames on average. Among all these sequences, there are 3,248 distractors,

which are distributing samples from false detection and tracking, increasing the difficulty of re-identification significantly. The 1,261 identities are captured by six non-overlapping cameras in this dataset, and each identity shows under at least two cameras. The dataset is split into two non-overlapping train and test sets, containing 625 and 636 identities, respectively.

Implementation Details. We re-implement the STA model as a baseline in our experiments. The input length of a video clip is set to $N = 4$ frames, extracted by random sampling of the original sequence. The number of horizontal parts of a feature map is set to $M = 4$. The input size of an image is 256×128. ResNet50 [12] pretrained on ImageNet [6] is used as the backbone network. The last spatial down-sampling operation of ResNet50 is removed by setting the stride of the last residual block to 1. Thus, the spatial resolution of feature maps is 16×8, with 2048 channels. The channel number of FC1 layers is 512, which leads to a 512-dimensional feature vector. For the training phase, following the common practices [20], random cropping and random horizontal flipping are used to augment the training data. We train this network on one single NVIDIA RTX 2080Ti GPU. Due to the limitation of its memory, we randomly sample 8 identities and 4 clips for each identity to organize a mini-batch with size of 32 for training, which is half of the size 64 in original paper. This might decrease the performance of baseline model a little, while the effectiveness of our proposed method could still be evaluated. For triplet loss, the margin parameter is set to 0.3 as recommended. All the other hyperparameters remain the same with [9]. Note that the baseline model in this paper is the STA model without extra feature fusion and inter-frame regularization strategies, so as to keep the baseline network vanilla.

Evaluation Protocol. In our experiments, we employ Cumulative Matching Characteristics (CMC) curve and mean Average Precision (mAP) as standard evaluation metrics. For CMC curve, we report its value at rank-1, rank-5 and rank-20, respectively.

4.2 Ablation Study

In this paper, we conduct two analytic experiments, namely ablation study, for average aggregation strategy (AAS) and raw feature utilization (RFU). This ablation study could evaluate the effectiveness of these two components in our proposed method. The experimental results are shown in Table 1 and Table 2, respectively.

Analysis of AAS. For average aggregation strategy, we carry out experiments to investigate the influence of sampled clip amount C and cosine similarity of final feature vector. The experimental results in Table 1 verify the effectiveness of our average aggregation strategy. In the first column of Table 1, *All* means that all the sampled clips cover the whole video sequence. It shows that the utilization of multiple clips improve the performance of model significantly, e.g., rank-1 accuracy from 83.5% to 85.7%, mAP from 77.5% to 80.0%. It is intuitive that using all the possible clips to compute an average of all the clip-level features, could generate a more robust feature representation of the whole

Table 1. Results with different average aggregation strategies.

Number of Clips	Cosine Similarity	rank-1	rank-5	rank-20	mAP
1	No	83.5	93.8	96.8	77.5
4	No	85.3	94.8	97.1	79.6
All	No	85.7	94.8	97.2	80.0
4	Yes	85.9	94.6	96.9	80.3
All	Yes	86.5	94.7	96.9	80.9

sequence. Moreover, compared to Euclidean distance (marked as *No* in the second column of Table 1), using cosine similarity to find the best match based on these averaged final features also improve both rank-1 and mAP, leading to 86.5% rank-1 and 80.9% mAP. We believe the reason behind this improvement is that the elements' magnitudes might be not as robust as the vector's angle in a high-dimensional space. Overall, this proposed average aggregation strategy improves the rank-1 by 3.0% and mAP by 3.4%.

Table 2. Ablation study of raw feature utilization.

Inference	Training	rank-1	rank-5	rank-20	mAP
No	No	86.5	94.7	98.2	80.9
Yes	No	87.2	95.8	97.1	82.3
Yes	Yes	87.7	96.4	98.3	82.3

Effectiveness of RFU. The experimental results of Table 1 are based on the feature vectors after the batch normalization, instead of the raw features. Here, we analyze the effect of raw feature utilization for video-based person re-identification. In our experiments, we perform inference using raw features (marked as *Yes* in the first column of Table 2), and then introduce these raw features to train a new model (marked as *Yes* in the second column of Table 2). The experimental results in Table 2 indicate that the utilization of raw features contributes to both inference and training stages. By removing the last batch normalization, the features with original distribution could not only help improve the feature aggregation of inference state but also training using triplet loss. Finally, it improves rank-1 by 1.2% and mAP by 1.4%.

4.3 Comparison with State-of-the-art

We compare our method to the state-of-the-art methods on the large-scale MARS dataset, including SeeForest [35], ReRank [34], RQEN [21], Diversity [15], RL [30], KHAN [22], Snippet-Sim [2] and STA [9]. Here, we report the result of STA, which is the full version with feature fusion strategy and inter-frame regularization. Table 3 shows our method is superior to the existing state-of-the-art, e.g., reaching rank-1 accuracy 87.7% and mAP 82.3%. In addition, we also

Table 3. Performance comparison on MARS dataset.

Method	rank-1	rank-5	rank-20	mAP
SeeForest [35]	70.6	90.0	97.6	50.7
ReRank [34]	73.9	-	-	68.5
RQEN [21]	77.8	88.8	94.3	71.1
Diversity [15]	82.3	-	-	65.8
RL [30]	83.1	91.3	-	69.9
KHAN [22]	85.7	94.3	97.2	77.8
Snippet-Sim [2]	86.3	94.7	98.2	76.1
STA [9]	86.3	95.7	98.1	80.8
STA + ReRank [9]	87.2	96.2	98.6	**87.7**
Ours	**87.7**	**96.4**	**98.3**	**82.3**
Ours + ReRank	**88.6**	**96.3**	**98.8**	87.4

report the results after re-ranking, where our method achieves 88.6% on rank-1 accuracy.

5 Conclusion

In this paper, a simple but efficient clip-level feature aggregation method for video-based person re-identification is proposed, which contains Average Aggregation Strategy (AAS) and Raw Feature Utilization (RFU). This parameter-free method can be applied to existing video-based person re-identification models which extract clip-level features, without extra layers or post-processing procedure. An ablation study is conducted to verify the effectiveness of AAS and RFU. Experimental results on MARS dataset demonstrate the improved accurary of our method. For the future work, an more intelligent way of selecting distinctive frames than random sampling from a long video sequence should be investigated. What is more, the combination of re-identification and tracking tasks is an interesting topic, which could lead to a better cross-camera tracking application.

Acknowledgments. The research leading to these results has received funding from the European Unions Horizon 2020 research and innovation programme under the Marie Skodowska-Curie grant agreement No 765866 - ACHIEVE.

References

1. Bazzani, L., Cristani, M., Perina, A., Farenzena, M., Murino, V.: Multiple-shot person re-identification by hpe signature. In: Proceedings of the IEEE International Conference on Pattern Recognition. pp. 1413–1416. IEEE (2010)
2. Chen, D., Li, H., Xiao, T., Yi, S., Wang, X.: Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1169–1178 (2018)

3. Cheng, D., Gong, Y., Zhou, S., Wang, J., Zheng, N.: Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1335–1344 (2016)
4. Cho, Y.J., Yoon, K.J.: Improving person re-identification via pose-aware multi-shot matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1354–1362 (2016)
5. Das, A., Chakraborty, A., Roy-Chowdhury, A.K.: Consistent re-identification in a camera network. In: Proceedings of the European Conference on Computer Vision. pp. 330–345. Springer (2014)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255. Ieee (2009)
7. Dimitrievski, M., Veelaert, P., Philips, W.: Behavioral pedestrian tracking using a camera and lidar sensors on a moving vehicle. Sensors **19**(2), 391 (2019)
8. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person re-identification by symmetry-driven accumulation of local features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2360–2367. IEEE (2010)
9. Fu, Y., Wang, X., Wei, Y., Huang, T.: Sta: Spatial-temporal attention for large-scale video-based person re-identification. In: Proceedings of the Association for the Advancement of Artificial Intelligence (2019)
10. Gheissari, N., Sebastian, T.B., Hartley, R.: Person reidentification using spatiotemporal appearance. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. vol. 2, pp. 1528–1535. IEEE (2006)
11. Gray, D., Tao, H.: Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: Proceedings of the European Conference on Computer Vision. pp. 262–275. Springer (2008)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
13. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 (2017)
14. Karaman, S., Bagdanov, A.D.: Identity inference: generalizing person re-identification scenarios. In: Proceedings of the European Conference on Computer Vision. pp. 443–452. Springer (2012)
15. Li, S., Bak, S., Carr, P., Wang, X.: Diversity regularized spatiotemporal attention for video-based person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 369–378 (2018)
16. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: Deep filter pairing neural network for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 152–159 (2014)
17. Liu, H., Jie, Z., Jayashree, K., Qi, M., Jiang, J., Yan, S., Feng, J.: Video-based person re-identification with accumulative motion context. IEEE Trans. Circuits Syst. Video Technol. **28**(10), 2788–2802 (2017)
18. Liu, K., Ma, B., Zhang, W., Huang, R.: A spatio-temporal appearance representation for viceo-based pedestrian re-identification. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3810–3818 (2015)
19. Luo, H., Gu, Y., Liao, X., Lai, S., Jiang, W.: Bag of tricks and a strong baseline for deep person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (2019)

20. McLaughlin, N., Martinez del Rincon, J., Miller, P.: Recurrent convolutional network for video-based person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1325–1334 (2016)
21. Song, G., Leng, B., Liu, Y., Hetang, C., Cai, S.: Region-based quality estimation network for large-scale person re-identification. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
22. Su, X., Qu, X., Zou, Z., Zhou, P., Wei, W., Wen, S., Hu, M.: k-reciprocal harmonious attention network for video-based person re-identification. IEEE Access **7**, 22457–22470 (2019)
23. 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 European Conference on Computer Vision. pp. 480–496 (2018)
24. Varior, R.R., Shuai, B., Lu, J., Xu, D., Wang, G.: A siamese long short-term memory architecture for human re-identification. In: Proceedings of the European Conference on Computer Vision. pp. 135–153. Springer (2016)
25. Wang, F., Xiang, X., Cheng, J., Yuille, A.L.: Normface: l_2 hypersphere embedding for face verification. In: Proceedings of the 25th ACM International Conference on Multimedia. pp. 1041–1049. ACM (2017)
26. Wang, T., Gong, S., Zhu, X., Wang, S.: Person re-identification by video ranking. In: Proceedings of the European Conference on Computer Vision. pp. 688–703. Springer (2014)
27. Wang, T., Gong, S., Zhu, X., Wang, S.: Person re-identification by discriminative selection in video ranking. IEEE Trans. Pattern Anal. Mach. Intell. **38**(12), 2501–2514 (2016)
28. Xiao, T., Li, S., Wang, B., Lin, L., Wang, X.: Joint detection and identification feature learning for person search. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3415–3424 (2017)
29. Yan, Y., Ni, B., Song, Z., Ma, C., Yan, Y., Yang, X.: Person re-identification via recurrent feature aggregation. In: Proceedings of the European Conference on Computer Vision. pp. 701–716. Springer (2016)
30. Zhang, W., He, X., Lu, W., Qiao, H., Li, Y.: Feature aggregation with reinforcement learning for video-based person re-identification. IEEE Trans. Neural Networks Learn. Syst. (2019). https://doi.org/10.1109/tnnls.2019.2899588
31. Zhao, R., Ouyang, W., Wang, X.: Person re-identification by salience matching. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2528–2535 (2013)
32. Zheng, L., Bie, Z., Sun, Y., Wang, J., Su, C., Wang, S., Tian, Q.: Mars: A video benchmark for large-scale person re-identification. In: Proceedings of the European Conference on Computer Vision. pp. 868–884. Springer (2016)
33. Zheng, L., Yang, Y., Hauptmann, A.G.: Person re-identification: Past, present and future. arXiv preprint arXiv:1610.02984 (2016)
34. Zhong, Z., Zheng, L., Cao, D., Li, S.: Re-ranking person re-identification with k-reciprocal encoding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1318–1327 (2017)
35. Zhou, Z., Huang, Y., Wang, W., Wang, L., Tan, T.: 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. pp. 4747–4756 (2017)