Large-Scale Video-Based Person Re-identification via Non-local Attention and Feature Erasing

Zhao Yang, Zhigang Chang, and Shibao Zheng(✉)

Institute of Image Communication and Network Engineering,
Shanghai Key Labs of Digital Media Processing and Transmission,
Shanghai Jiao Tong University, SEIEE Buildings 5-421, Shanghai 200240, China
{10110907,changzig,sbzh}@sjtu.edu.cn

Abstract. Encoding the video tracks of person to an aggregative representation is the key for video-based person re-identification (re-ID), where average pooling or RNN methods are typically used to aggregating frame-level features. However, It is still difficult to deal with the spatial misalignment caused by occlusion, posture changes and camera views. Inspired by the success of non-local block in video analysis, we use a non-local attention block as a spatial-temporal attention mechanism to handle the spatial-temporal misalignment problem. Moreover, partial occlusion is widely occurred in video sequences. We propose a local feature branch to tackle the partial occlusion problem by using feature erasing in the frame-level feature map. Therefore, our network is composed by two-branch, the global branch via non-local attention encoding the global feature and the local feature branch grasping the local feature. In evaluation, the global feature and local feature are concatenated to obtain a more discriminative feature. We conduct extensive experiments on two challenging datasets (MARS and iLIDS-VID). The experimental results demonstrate that our method is comparable with the state-of-the-art methods in these datasets.

Keywords: Video-based person re-ID · Non-local attention · Feature erasing

1 Introduction

Person re-Identification (re-ID) aims to match the same pedestrian across multiple non-overlapping camera at different places or time, which has received more and more attention due to its great significance to applications in recent years, such as video surveillance [25], activity analysis [17], tracking [28] and so on. Deep learning methods have shown to be effective for person re-identification and have made great progress than traditional approaches [1,4,7]. However, it is still a difficult and challenging problem because of the huge variations of pedestrian's

G. Zhai et al. (Eds.): IFTC 2019, CCIS 1181, pp. 327–339, 2020.
https://doi.org/10.1007/978-981-15-3341-9_27

appearance caused by person's postures, background clutter, camera viewpoints, blur and occlusion.

In general, with respect to the "probe-to-gallery" pattern, there are four person re-identification strategies: image-to-image, image-to-video, video-to-image, and video-to-video [30]. Image-based person re-identification has been widely studied in the literature where person are matched by only a single still image. It's obviously that video data contain much richer appearance information than a single image which is beneficial to identify a person under complex conditions including occlusion, blur, and the changes of camera viewpoint [31]. Therefore, we will concentrate on the video-based person re-identification in this paper, which associates the trajectories of the pedestrian by comparing his/her video sequences not only one still image.

For the Video-based person re-ID task, the key is to learn a function which encoding the video tracks of person to a single feature in a lower dimensional feature space. A typical pipeline of video-based person re-id methods contains a frame-level feature extractor (Convolutional Neural Network) and a feature aggregation module to aggregate the frame-level features to a single feature, where average pooling or maximum pooling is widely used. However, basic average pooling or maximum pooling is hard to handle spatial misalignment caused by the variation of human poses or viewpoints among sequences. To make full use of the video sequence information to alleviate the noise, many of the recent methods focus on the temporal model. There are two classical temporal modelling methods: Recurrent Neural Network (RNN) based and temporal attention based. In RNN-based methods, McLanghlin et al. [19] proposed to use an RNN to aggregate feature among the temporal frame-level features; Yan et al. [27] also used an RNN to encode sequence features, where the final hidden state is used as the output sequence representation. In temporal attention based methods, Liu et al. [16] proposed a Quality Aware Network (QAN), to measure the quality of every frame among the sequence and score. QAN is actually an attention weighted average of frame-level feature; Zhou et al. [31] proposed to encode the video sequence with temporal RNN and attention to pick out the most discriminative frames. However, these method mentioned above are still difficult to handle temporal attention and spatial misalignment simultaneously. Moreover, these method mainly consider the most discriminative frames in a video sequence by attention weights to learn a global feature representation, which ignored the discriminative body parts in a frame. Local discriminative body parts learning have a significant improvement to occlusion, particularly partial occlusion caused by other pedestrians or blackground cluster. Li et al. [14] proposed a spatial-temporal attention model that automatically discovers a diverse set of distinctive body parts.

In this paper, inspired by the success of non-local block in video classification [26], we propose a spatial-temporal attention mechanism to tackle the spatial misalignment problem by adopting a non-local block in a bottleneck structure. Moreover, We propose a local feature branch to tackle the partial occlusion problem by using unified region feature erasing in the frame-level fea-

ture map. Therefore, our network is composed by two-branch, the global branch via non-local attention scheme encoding the global feature and the local feature branch grasping the local feature. In evaluation, the global feature and local feature are concatenated to generate a more discriminative feature. In summary, our contribution of this work are three fold:

1. We propose a non-local block with a bottleneck structure as a spatial-temporal attention module to deal with the misalignment problem intra the video sequences, which could learn a more robust global feature.
2. The positional consistency feature erasing mechanism intra a video sequence in the local feature branch can effectively learn the local distinctive feature, which is beneficial to the partial occlusion problem.
3. A more discriminative feature used to re-identify the person is obtained by fusioning the robust global feature learned by the non-local attention and the local feature learned by feature erasing scheme with positional consistency. We conduct extensive experiments and ablation study to demonstrate the effectiveness of each component in a two challenging datasets: Mars and iLIDS-VID. The experiment results are comparable with the state-of-the-art methods in these datasets.

2 Related Works

In this section, we review the related work including image-based person re-identification, video-based person re-identification and self-attention.

2.1 Image-Based Person Re-identification

Image-based person re-identification has been widely studied in the literature. Recent work focus on discriminative feature learning and metric learning. Hermans et al. [12] proposed a variant of triplet loss to perform end-to-end deep metric learning, which selects the hardest positive and negative example for every anchor example in a batch and outperforms many other published methods by a large margin. Recently many work try to learn fine-grained discriminative features from local part of person. Part-based methods [9,20,23] have achieved state-of-the-art results, which split the feature map horizontally into a fixed number of strips and aggregate features from those strips. However, in evaluation, the high-dimensional feature vector increase the burden of computation and storage. Dai et al. [5] proposed a batch feature erasing method which erasing the feature map of images from different person in the batch at the same position can effectively learn the local feature without much computation. Inspired by this method, we extend it to our video-based person re-identification network to learn local discriminative feature of the video sequence. The feature maps of every frame intra the sequence are synchronously erased in the same position in a batch.

2.2 Video-Based Person Re-identification

Video-based person re-identification is an extension of the image-based person re-identification and has drawing more and more attraction due to the richer information contained in a video sequence. Making full use of the temporal feature of the video sequence is the most important part of video-based person re-identification. Most of the recent methods focus on the design of spatial-temporal attention intra the sequence. Gao et al. [10] revisited many temporal modeling method, including rnn-based, average pooling and proposed a attention scheme by using a fully connected layer to weighted every frame. Distinct with these methods which represent the video sequence in a single feature, Chen et al. [3] devided the video sequence to several overlapped snippets, and then computed the similarity of every snippet of two video sequences with co-attentive embedding and assumed the average scores of the top 20% as the final similarity of the two video sequences, which can implicitly deal with the posture misalignment problem.

2.3 Self-attention

Long-term range dependence modelling makes a great significance to natural language process and computer vision. Non-local means [2] is a classical filtering algorithm in digital image process that computes a weighted average of all pixels in an image. Different from convolution operation which only consider the local neighborhood pixels, non-local means allows distant pixels to contribute to the filtered response at a location based on patch appearance similarity. Vaswani et al. [22] proposed a self-attention module which computes the response at a position (e.g., a word) in a sequence (e.g., a sentence) by taking every position into account and summing weighted responses in an embedding space. Furthermore, Wang et al. [26] proposed a non-local module to bridge self-attention for machine translation to the more general class of non-local filtering operations that are applicable to image and video problems in computer vision. The non-local module has the ability to model the long-term range dependence and could discover the consistent semantice messages of frames in a video sequence. Therefore, we can embed a non-local block to our neural network to capture a long-term range dependence in video both on space and time to deal with the spatial-temporal misalignment problem intra the sequence.

3 The Proposed Method

In this section, we introduce the overall architecture (Fig. 1) of the proposed method, and then describe each of its important part with more details. First the input video is devided into several non-overlapped consecutive clips $\{c_k\}$, each contains T frames. The backbone CNN extracts feature maps for every frame in the clip, and then the non-local block takes these feature maps as input to model the spatial-temporal attention intra the clip, and obtain the weighted

feature maps without dimension change of the input feature maps. The rest part have two component: the global feature branch and the local feature branch. The global one uses global average pooling over the weighted feature maps to grasp the global feature, and the local branch adopts feature erasing by a mask intra the clip with position consistency to learn the local discriminative feature. In evaluation, we concatenate the global feature and the local feature to generate a more distinctive and robust clip-level feature. The final video-level feature is the average of all clip-level feature. For the loss function, we combine batch hard triplet loss and softmax loss [12].

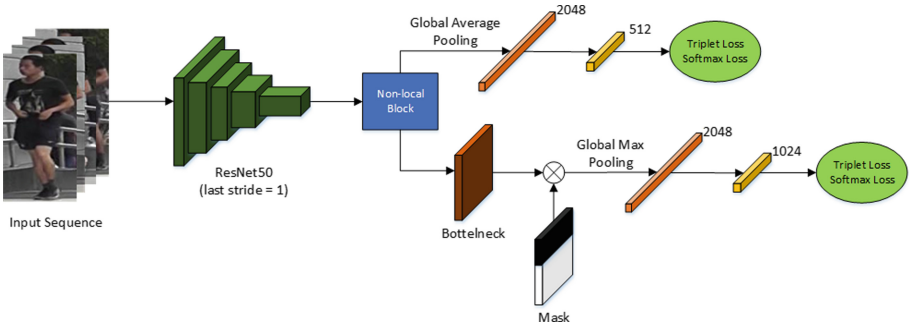

Fig. 1. Overall architecture. The input video clip is fed into the backbone network to be transformed to the feature maps, then the non-local block captures the long range dependence intra the video clip both in space and time, and output weighted feature maps. The global branch adopts global average pooling to the weighted feature maps to generate a global feature representation, and the local branch uses a feature erasing mask to grasp the discriminative local feature. In evaluation, we concatenate the 512 dimension global feature vector (L2 normalized) and the 1024 dimension local feature vector (L2 normalized) to generate the final robust and discriminative representation of the input clip.

3.1 Backbone Network

We adopt ResNet50 [11] as the backbone network. Sun et al. [20] removed the last spatial downsampling in the last residual block in ResNet50 to increase the size of feature map. Luo et al. [18] denoted the last spatial down-sampling operation in the ResNet50 backbone as last stride. Thus, we modify the last stride of ResNet50 from original 2 to 1. For a classical input image size 256×128, the modified ResNet50 outputs a feature map with the spatial size of 16×8 and channel dimension 2048.

3.2 Non-local Block

A non-local attention block has the ability to capture the long-term range dependency in sequence or video both in space and time. Therefore, for video-based

person re-identification, a non-local attention block could tackle the spatial misalignment problem caused by the viewpoint and distance. In this part, we give an detail definition of non-local operation.

Following the non-local mean operation [2,26], a generic non-local operation in deep neural networks is defined as:

$$y_i = \frac{1}{C(x)} \sum_{\forall j} f(x_i, x_j) g(x_j) \tag{1}$$

Here x_i can be the feature vector of the input signal x (image, sequence, video; often their features) at the corresponding position i. y is the output signal of the same size as x. A pairwise function f calculate a scalar which represent the relationship such as similarity between x_i and all x_j The unary function g computes a representation of x_j in an embedded space. Finally, the response y_j is normalized by a factor $C(x)$.

There are several versions of f and g, e.g. gaussian, dot product, embedded gaussian. In our experiment, we select embedded gaussian version of non-local, which is the essential principle of self-attention module presented for machine translation [26]. Embedded gaussian function f is given by:

$$f(x_i, x_j) = e^{\theta(x_i)^T \phi(x_j)} \tag{2}$$

Here $\theta(x_i) = W_\theta x_i$ and $\phi(x_j) = W_\phi x_j$. If we set $C(x)$ as a softmax function, we would have:

$$y = softmax(x^T {W_\theta}^T W_\phi x) g(x) \tag{3}$$

A non-local operation is a flexible building block and can be easily inserted to neural network. In the non-local neural network [26], a non-local block is defined as:

$$z_i = W_z y_i + x_i \tag{4}$$

where y_i is defined in Eq. (1) and "$+x_i$" represent a residual connection [26]. The residual connection allows us to insert a new non-local block into any pre-trained model, without breaking its initial behavior (e.g., if W_z is initialized as zero) [11]. An example non-local block is illustrated in Fig. 2. In our experiment, we insert a non-local block within a bottleneck structure [11], which sequentially contains a conv1×1, a conv3×3, a conv1×1, each followed by a batchnorm layer and ReLU function, the non-local block is inserted between the last batchnorm layer and last ReLU function.

3.3 Unified Region Feature Erasing

We extend the batch feature erasing method [5] proposed in image-based person re-id to video-based, which we defined as unified region feature erasing. In a batch of video clips, especially intra every clip, the region erased is the same, which is beneficial to grasp the same pattern local feature in a batch. The region erased should be big enough to include part semantice messages of the original feature map. The detailed algorithm is described below:

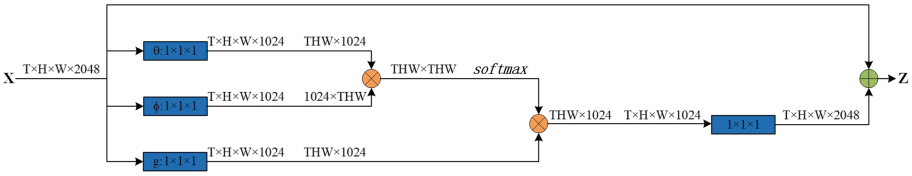

Fig. 2. A spacetime non-local block. The feature maps are shown as the shape of their tensors, e.g., $T \times H \times W \times 2048$ for 2048 channels (proper reshaping is performed when noted). "$\otimes$" denotes matrix multiplication, and "$\oplus$" denotes element-wise sum. The softmax operation is performed on each row. The blue boxes denote $1 \times 1 \times 1$ convolutions. Here we show the embedded Gaussian version, with a bottleneck of 1024 channels.

Algorithm 1. Unified Region Feature Erasing Algorithm.

Require:
Input tensor of size [N, T, C, H, W], T_{in};
Ratio of erased height, r_h;
Ratio of erased width, r_w;
Ensure:
Erased Tensor, T_e;
if training **then**
  $H_e := H \times r_h, W_e := W \times r_w$
  $x_e := Rand(0, H - H_e), y_e := Rand(0, W - W_e)$
  $Mask := Ones(H, W)$
  $Mask[x_e : x_e + H_e, y_e : y_e + W_e] := 0$
  $T_e := T_{in} \times Mask$
else
  $T_e := T_{in}$
end if
return T_e;

3.4 Loss Functions

We use a triplet loss function and a softmax cross-entropy loss function with label smoothing regularization [21] to train the network. We use the batch hard triplet loss function which was originally proposed in [12]. To form a batch, we randomly sample P identities and randomly sample K clips for each identity (each clip contains T frames); Totally there are PK clips in a batch. For each sample a in the batch, the hardest positive and the hardest negative samples within the batch are selected when forming the triplets for computing the loss $L_{triplet}$.

$$L_{tripletloss} = \sum_{i=1}^{P}\sum_{a=1}^{K}[m + \overbrace{\max_{p=1\ldots k} D(f_a^i, f_p^i)}^{hardest\ positive} - \underbrace{\min_{\substack{j=1\ldots P \\ n=1\ldots K \\ j!=i}} D(f_a^i, f_n^j)}_{hardest\ negative}] \tag{5}$$

The softmax cross-entropy loss function with label smoothing regularization is given by:

$$L_{softmax}^{smooth} = -\frac{1}{P \times K}\sum_{i=1}^{P}\sum_{a=1}^{K} p_{i,a} log((1-\epsilon)q_{i,a} + \frac{\epsilon}{N}) \tag{6}$$

where $p_{i,a}$ is the ground truth identity, $q_{i,a}$ is the prediction of sample $\{i, a\}$ and N is the number of classes. The final loss function is the combination of the two losses:

$$L_{total} = L_{tripletloss} + L_{softmax}^{smooth} \tag{7}$$

4 Experiments

4.1 Datasets and Evaluation Protocal

Mars Dataset. [30] is one of the largest video-based person re-identification dataset. It contains 17,503 tracklets from 1,261 identities, and additional 3,248 tracklets serving as distractors. Each tracklet has 59 frames on average. These video tracklets are captured by six cameras in a university campus. The total 1,261 identities are split into 625 identities for training and 636 identities for testing. The ground truth bounding boxes are detected and tracked using the Deformable Part Model (DPM) [8]and GMCP tracker [29].

iLIDS-VID Dataset. [24] consists of 600 video sequences of 300 persons. Each image sequence has a variable length ranging from 23 to 192 frames, with averaged number of 73. This dataset is challenging due to clothing similarities among people and random occlusion. The ground truth bounding boxes are annotated manually.

Evaluation Protocol. In our experiments, we report the standard evaluation metrics: mean average precision score (MAP) and the cumulative matching curve (CMC) at rank-1, rank-5, rank-10 and rank-20. For fair comparison, we report MAP and CMC for Mars dataset, CMC for iLIDS-VID dataset.

4.2 Implemantation Details

As mentioned in "The Proposed Method", we use modified ResNet50 which is pre-trained on the ImageNet [6]. The frame number of each clip is set to $T = 4$. we augment the clip data with clip-level random synchronize flipping and cropping. There are $P \times K$ clips in a minibatch, we set $K = 4$ (clips of each person), $P = 32$ (identities) and $K = 4$, $P = 16$ for Mars and iLIDS-VID respectively, where all of P and K are randomly sampled. The input image is resized to 256×128. We set $r_w = 1.0$ and $r_h = 1/3$ to erase the feature map. The margin parameter in triplet loss is equal to 0.3. During training, we use the Adam optimizer [13] with weight decay 5e−4. We use a warm up training scheme, the initial learning rate is linearly increasing from 1e−4 to 1e−3 in first 50 epochs. And the learning rate decays to 1e−4, 1e−5, 1e−6 at 200, 400, 600 epochs respectively. We train the network for 800 epochs in total.

4.3 Ablation Study

We conduct ablation investigation to analyze the effect of each component in our proposed method, including non-local block and the local feature erasing branch.

Effectiveness of Components. In Table 1, we show the results of each component. All these three methods adopt the modified ResNet50 as the backbone network, and use batch hard triplet loss and softmax cross-entropy loss function with label smoothing regularization for a fair comparison. *Baseline* method represents our global branch without the non-local block trained on MARS/iLIDS-VID dataset, where the global average pooling layer encoding the feature maps of clip to a single tensor with dimension 2048. Then the feature dimension is reduced from 2048 to 512 by using a fully connected layer. We choose the final 512 dimension feature vector to represent the clip. $Baseline + NL$ method represents our global branch with the non-local block. Compared to *Baseline*, the $Baseline + NL$ method improves rank-1 and mAP accuracy by 1.0% and 1.4% on the large-scale Mars dataset, as well as 6.6% and 4.4% on iLIDS-VID dataset. The result demonstrate that the non-local block is effectively to model the spatial-temporal attention to learn a more discriminative feature and deal with the spatial misalignment problem. The visualization results are show in Fig. 3. The $Baseline + NL + FE$ method is our whole network with a non-local block to model the long range dependency within a video both space and time, the final feature representation is the concatenation of the feature vector from the global branch and the feature vector from the local branch. Our overall network increase 3.2% and 3.5% in rank-1 and mAP when compared with *Baseline* on Mars dataset, and gain 8.6% and 5.9% improvements on iLIDS-VID dataset. Obviously, combining the global feature and local feature together can effectively obtain a more robust and discriminative feature to re-identify the same person.

Table 1. Comparison of different proposed components.

Model	MARS					ILIDS-VID				
	R1	R5	R10	R20	MAP	R1	R5	R10	R20	MAP
Baseline	83.5	94.3	96.5	97.6	77.5	76.7	94.0	97.0	100.0	84.8
Baseline + NL	84.5	94.9	96.6	97.7	78.9	83.3	96.7	99.3	100.0	89.2
Baseline + NL + FE	86.7	95.4	97.2	97.9	81.0	85.3	98.0	99.3	99.3	90.7

4.4 Comparison with the State-of-the-Arts

Tables 2 and 3 report the comparison of our proposed method with the state-of-the-art methods on Mars and iLIDS-VID respectively.

Fig. 3. Visualization of the behavior of a non-local block to deal with spatial misalignment.

Results on MARS. Mars is one of the largest video-based person re-id dataset and full of challenge. Table 2 show comparisons between our proposed method with most of the state-of-the-art method on Mars dataset. We achieve 81.0% in MAP, which outperforms all these work by nearly 5%. For the rank-1, rank-5, and rank-10, Our approach all keeps ahead than all these methods. Our approach achieve 97.9% in rank-20, which is comparable with the $Snipped + OF^*$. $Snipped + OF^*$ uses optical flow to provide more extra motion information and brings with more computation.

Table 2. Comparison of our proposed method with the state-of-the-art on MARS dataset. '–': no reported results.

Model	Rank1	Rank5	Rank10	Rank20	MAP
Mars [30]	68.3	82.6	–	89.4	49.3
SeeForest [31]	70.6	90.0	–	97.6	50.7
QAN [16]	73.7	84.9	–	91.6	51.7
Non-local + C3D [15]	84.3	94.6	96.2	–	77.0
STAN [14]	82.3	–	–	–	65.8
Snipped [3]	81.2	92.1	–	–	69.4
Snipped + OF^* [3]	86.3	94.7	–	**98.2**	76.1
Our proposed	**86.7**	**95.4**	**97.2**	97.9	**81.0**

Results on iLIDS-VID. iLIDS-VID is a small dataset especially compared to Mars dataset. Table 3 show comparisons between our proposed method with most of the state-of-the-art method on iLIDS-VID dataset. Our method surpass most of these method, and is comparable with the best method $Snipped + OF^*$. For rank-5 and rank-10, we achieve the best results with some advantages.

Table 3. Comparison of our proposed method with the state-of-the-art on iLIDS-VID dataset. '–': no reported results.

Model	Rank1	Rank5	Rank10	Rank20
Mars [30]	53.0	81.4	–	–
SeeForest [31]	55.2	86.5	–	97.0
QAN [16]	68.0	86.8	95.4	97.4
STAN [14]	80.2	–	–	
Snipped [3]	79.8	91.8	–	–
Snipped + OF^* [3]	**85.4**	96.7	98.8	**99.5**
Our proposed	85.3	**98.0**	**99.3**	99.3

5 Conclusion

This paper concentrates on the large-scale video-based person re-identification. The key of video-based person re-id is to learn a mapping which encoding the clip-level feature to a single feature in a low-dimensional feature space. We adopt the non-local block to capture the long-term range dependence in a video both space and time, which can learn the corresponding consistent semantic information to deal with the spatial misalignment occurred in video sequences. The synchronous feature erasing scheme is beneficial to learn the local discriminative feature, which can alleviate the partial occlusion problem intra the video sequence. A more robust feature is generated by concating the global feature and local feature. Extensive experiments conducted on two challenging datasets (Mars and iLIDS-VID) demonstrate the effect of our method, which is comparable to most of state-of-the-art methods. A valuable direction of person re-identification including image and video is to combine with multi-pedestrian tracking in real world.

References

1. Ahmed, E., Jones, M., Marks, T.K.: An improved deep learning architecture for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3908–3916 (2015)
2. Buades, A., Coll, B., Morel, J.M.: A non-local algorithm for image denoising. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), vol. 2, pp. 60–65. IEEE (2005)
3. 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)
4. Chen, S.Z., Guo, C.C., Lai, J.H.: Deep ranking for person re-identification via joint representation learning. IEEE Trans. Image Process. **25**(5), 2353–2367 (2016)
5. Dai, Z., Chen, M., Zhu, S., Tan, P.: Batch feature erasing for person re-identification and beyond. arXiv preprint arXiv:1811.07130 (2018)

6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
7. Ding, S., Lin, L., Wang, G., Chao, H.: Deep feature learning with relative distance comparison for person re-identification. Pattern Recogn. **48**(10), 2993–3003 (2015)
8. Felzenszwalb, P.F., McAllester, D.A., Ramanan, D., et al.: A discriminatively trained, multiscale, deformable part model. In: CVPR, vol. 2, p. 7 (2008)
9. Fu, Y., et al.: Horizontal pyramid matching for person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 8295–8302 (2019)
10. Gao, J., Nevatia, R.: Revisiting temporal modeling for video-based person ReID. arXiv preprint arXiv:1805.02104 (2018)
11. 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)
12. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 (2017)
13. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
14. 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)
15. Liao, X., He, L., Yang, Z., Zhang, C.: Video-based person re-identification via 3d convolutional networks and non-local attention. In: Jawahar, C.V., Li, H., Mori, G., Schindler, K. (eds.) ACCV 2018. LNCS, vol. 11366. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-20876-9_39
16. 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, pp. 5790–5799 (2017)
17. Loy, C.C., Xiang, T., Gong, S.: Multi-camera activity correlation analysis. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1988–1995. IEEE (2009)
18. 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)
19. 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)
20. 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 (ECCV), pp. 480–496 (2018)
21. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826 (2016)
22. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
23. Wang, G., Yuan, Y., Chen, X., Li, J., Zhou, X.: Learning discriminative features with multiple granularities for person re-identification. In: 2018 ACM Multimedia Conference on Multimedia Conference, pp. 274–282. ACM (2018)

24. Wang, T., Gong, S., Zhu, X., Wang, S.: Person re-identification by video ranking. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8692. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10593-2_45
25. Wang, X.: Intelligent multi-camera video surveillance: a review. Pattern Recogn. Lett. **34**(1), 3–19 (2013)
26. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7794–7803 (2018)
27. Yan, Y., Ni, B., Song, Z., Ma, C., Yan, Y., Yang, X.: Person re-identification via recurrent feature aggregation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9910. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46466-4_42
28. Yu, S.I., Yang, Y., Hauptmann, A.: Harry Potter's Marauder's map: localizing and tracking multiple persons-of-interest by nonnegative discretization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3714–3720 (2013)
29. Roshan Zamir, A., Dehghan, A., Shah, M.: GMCP-Tracker: global multi-object tracking using generalized minimum clique graphs. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7573. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33709-3_25
30. Zheng, L., et al.: MARS: a video benchmark for large-scale person re-identification. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9910. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46466-4_52
31. 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)