Intra-clip Aggregation for Video Person Re-identification

Takashi Isobe  
jbj18@mails.tsinghua.edu.cn

Jian Han  
jianhan.thu@gmail.com

Fang Zhu  
fz723@nyu.edu

Yali Li  
liyali@tsinghua.edu.cn

Shengjin Wang  
wgsj@tsinghua.edu.cn

1 Tsinghua University
2 New York University

Abstract

Video-based person re-id has drawn much attention in recent years due to its prospective applications in video surveillance. Most existing methods concentrate on how to represent discriminative clip-level features. Moreover, clip-level data augmentation is also important, especially for temporal aggregation task. Inconsistent intra-clip augmentation will collapse inter-frame alignment, thus bringing in additional noise. To tackle the above-mentioned problems, we design a novel framework for video-based person re-id, which consists of two main modules: Synchronized Transformation (ST) and Intra-clip Aggregation (ICA). The former module augments intra-clip frames with the same probability and the same operation, while the latter leverages two-level intra-clip encoding to generate more discriminative clip-level features. To confirm the advantage of synchronized transformation, we conduct ablation study with different synchronized transformation scheme. We also perform cross-dataset experiment to better understand the generality of our method. Extensive experiments on three benchmark datasets demonstrate that our framework outperforming the most of recent state-of-the-art methods.

1 Introduction

Person re-identification (Re-id) aims to recognize the same person in different images or videos captured by different cameras distributed at separated physical locations. With the emergence of deep learning methods within recent years and their influence on the computer domain, video-based person re-id is in high demand for video surveillance and has driven significant progress. Specifically, video-based person re-id is a generalized image-based person re-id problem. Given a query video of one person and a gallery set which contains a number of candidate videos, the system aims to compute the clip-level distance metric and find the query person.
In contrast with image-based person re-id which is consequently limited by the quality of image [5, 20, 25, 27], video-based person re-id is more robust to noise. Diversity information between consecutive frames can be adopted to complement a single-frame limited representation [17, 24, 35], especially the frame is corrupted by occlusion or lighting. Most earlier work [16, 19, 22, 35] concentrate on taking full use of inter-clip temporal information to alleviate the spatial noise. Unfortunately, in those methods, extracting temporal information between two consecutive frames need much more parameters and lead to an excessively heavy computational complexity. To effective harness the multi-frame information while maintains low computational cost, temporal pooling has been widely used in most existing method as the temporal aggregation scheme, which represent each frame as feature vector and then aggregate them across time using average or maximum pooling [5, 6, 15]. However, directly average of max the frame-level features may destroy the dependency between each feature, which means original feature domains has been changed. Moreover, temporal pooling is a non-parameter operation, which not able to learn rich information from the new domain. Thus the aggregated clip-level features may not be sufficiently discriminative for identifying different persons. In this paper, we propose a Intra-clip Aggregation (ICA) module that aggregate the frame-level features and then use a group of convolution kernels to learn the distinctive clip-level features. Instead of only encoding the whole frame once time, we concatenate a learnable module after temporal pooling operation in order to learn much more information from clip-level feature. To maintain relative fewer parameters, we design the learnable module with a bottleneck structure, following ResNet [9]. Due to the over-fitting trap, we conduct cross-dataset experiment to prove the generality of our model. Data augmentation is a crucial image-wise initialization technique as well as parameter-wise initialization in networks [7, 8]. In particular, video-based person re-id is essential to clip-level distance metric learning tasks. Handling a sequence of images inconsistently will collapse inter-frame alignment, thus bring in additional noise. To our knowledge, the prior work [6, 15, 24] employ image-level data augmentation techniques to randomly augmented each frame (e.g. cropping). In order to deal with the unreasonable data augmentation in video-based person re-id, we synchronically augments the inter-clip frames with cropping, flipping as well as erasing [14, 34].

To sum up, the main contributions of our work are in two-folds: (I) Due to the deficient temporal aggregation scheme and unreasonable data augmentation, we design a novel framework for video-based person re-id, which consists of two main modules to tackle with the above-mentioned problem: Intra-clip Aggregation (ICA) and Synchronized Transformation (ST). The former module leverages two-level intra-clip encoding to generate more distinctive clip-level features, while the latter synchronously augments intra-clip frames. Consequently, our temporal aggregation scheme with reasonable clip-level data augmentation can generate more discriminative clip-level feature, while maintains relatively low computational costs. (II) We perform extensive experiments on two benchmark datasets, ILIDS-VID [28] and MARS [33]. Our method outperforms state-of-the-art methods without re-ranking. To evaluate the generality of our method, we conduct cross-dataset experiment on PRID2011 [11].

The rest part of this paper is organized as follows, We first give an overview of related work in Sec 2, then we introduce the overall system framework for video-based re-id in Sec 3, and describe the proposed approach in details. In Sec 4, we conduct and analyze the proposed method by extensive experiments and comparisons with other methods.
2 Related Work

In this section, we give an overview about related work, including distance metric learning for person re-id, temporal aggregation, and data augmentation techniques.

**Distance Metric Learning for Person Re-id.** Most previous work on image-based person re-id focused on designing discriminative features [5, 12, 20] and robust distance metrics [1, 4, 30]. Recently, by combining impactful human semantic information [25, 27] and robust loss function [10, 29] under a deep Convolutional Neural Network (CNN) [9], image-based person re-id has achieved impressive progress. As for distance metric learning, Wojke et al. [30] presented a Cosine SoftMax Classifier which enforces a cosine similarity on the representation space. Hermans et al. [10] proposed the triplet loss to relax the contrastive formulation allowing samples to move more freely as long as the margin is kept constant. In [1], the authors split the metric into texture and color components in order to learn deep color-invariant features and color difference between a pair of cameras respectively. Metric learning is also a challenging problem for video-based person re-id. You et al. [32] developed a top-push metric learning method to minimize intra-class variations and maximize inter-class distance.

**Temporal Aggregation.** Recently, video re-identification has drawn significant attention. Learning a discriminative clip-level feature is crucial to video-based person re-id. Most previous work dedicated to aggregating frame features vectors across temporal dimension into a clip-level feature [6, 16, 17, 18, 35]. In [35], the authors extracts and aggregates the temporal and spatial information between consecutive frames simultaneously with one-stream Neural Network. However, in this way, the spatial feature of most frame are repeatedly learned, hence which are susceptible to frame appearance. To overcome the limitation of one-stream framework, [17] exploit additional network branch to extract optical flows between consecutive frames as the temporal information. Thereon, separately extract the spatial and temporal features from RGB images and optical flow by two independent but weight sharing CNN branch, and then aggregate them in intermediate layer. Liao et al. [16] employed a succession of 3D convolution kernel pre-trained on kinetics to extract spatial and temporal features simultaneously from a video volume, which keeps the intra-clip consistency and learns the context information of local appearance patch. On the other hand, temporal alignment is a key point to temporal pooling performance [15, 18, 24]. Li et al. [15] create a compact encoding of the video that exploits useful partial information in each frame. Liu et al. [18] adopted historical appearance and motion context to search the missing parts and suppress noisy parts. Song et al. [24] employ the high-quality region, which is predicted by region-based quality predictor part, to compensate the influence of an image region with poor quality.

**Data Augmentation.** Data augmentation, an explicit form of regularization, has been widely used in the training process of CNN. Generally, this technique enlarge the training set by various transformation, in order to best improve the classifier. Such as, cropping, flipping, normalization, eliminating, colour shifting etc. [14, 23, 34]. So far, it has witnessed great success not only in classifying and detection but also in image-based person re-identification. Simonyan et al. [23] enlarged the original dataset by adjusting the attribute of brightness, saturation and contrast of image probability. Zhong et al. [34] designed a new data augmentation technique which randomly occludes an arbitrary region of the input image during each training iteration. In this paper, we augments clip-level data with synchronized cropping, flipping and erasing.
3 Methodology

Approach Overview. We aim to enable the network to learn more discriminative and robust clip-level features with relatively fewer parameters and lower computational cost. Temporal pooling is an effective temporal aggregation scheme, which averages or maximizes the different video frames to resist spatial noise. However, the composite features derived from temporal pooling lack of discriminative representativeness due to the fact that the temporal pooling cannot learn from clip-level feature. To handle this problem, we proposed an Intra-clip Aggregation (ICA) module that aggregates the frame-level features and then uses a group of convolution kernels to learn the distinctive clip-level features. In particular, compared with simple temporal pooling baseline, our method generates more discriminative clip-level features while maintains fewer computational load. We will introduce more details about ICA module in Sec.3.2. The other contribution of our method is clip-level data augmentation techniques. Video frame appearance plays an important factor in temporal pooling methods. However, frame-based data augmentation techniques randomly augmented each frame that make the frame unaligned and bring in additional noise factor for each frame. To address this problem, we come up with clip-level initialization techniques, which synchronously augment intra-clip frames. Overall, our person re-id matching model consists of two iterative procedure: (1) synchronized data augmentation and (2) generalized clip-level feature aggregation, as elaborated in the following.

3.1 Synchronized Data Augmentation Based on Video Re-id

In this section, we describe the temporal data augmentation (temporal Transform) techniques in details. Our scheme expands the image-level augmentation techniques to clip-level, which uniformly transforms all frame of a given sequence, just like artificially cropping or erasing the same patch of each frame. We formulate the random transformation and synchronized transformation as following.

\[
T_{rt}(f_1^c, f_2^c, \ldots, f_T^c) = \{\Psi_3(\Psi_2 | \Psi_1(f_1^c)), \Psi_3(\Psi_2 | \Psi_1(f_2^c)), \ldots, \Psi_3(\Psi_2 | \Psi_1(f_T^c))\}
\]

(1)

\[
T_{st}(f_1^c, f_2^c, \ldots, f_T^c) = \{\Upsilon_3(\Upsilon_2 | \Upsilon_1(f_1^c, f_2^c, \ldots, f_T^c))\}
\]

(2)

Where \(T_{rt}\) and \(T_{st}\) is the operation of image augmentation and temporal augmentation. \(\{f_k^c\}_{k=1}^T\) denotes the set of frames of a given sequence. \(\{\Psi_3, \Psi_2, \Psi_1\}\) and \(\{\Upsilon_3, \Upsilon_2, \Upsilon_1\}\) represent three type of image-level and clip-level data augmentation techniques respectively. As illuminated in Figure 1, we implement temporal transform on data processing layer, which incorporates three-part synchronized augmentation techniques: cropping, flipping and erasing. To further analyze the performance of temporal augmentation, we conduct exploratory experiment on Sec.4.2 and present a group of visualization result to support our analysis.

3.2 Intra-clip Aggregation Module

In this section, we describe the structure details of the proposed Intra-clip Aggregation (ICA) module. Figure 2 shows that our ICA module consists of two sub-modules. The left part exploit two sequential non-parameter encoding operation to generate a temporary clip-level features, and another part utilize a clip-level encoding operation to reconstruct the clip-level
Figure 1: Overall system architecture consists of two mainly transformation module: (1) Synchronized Transformation and (2) ICA module. The left part illuminates our person re-id pipeline, which has three important parts: a clip-level data transformer, an image-level feature extractor and a clip-level feature generator. The right part shows the processing pipeline of each module (A: Synchronized Cropping, B: Synchronized Erasing and C: Overall Transformation).

We formulate the intra-clip pooling operation as following:

\[ Y_{1,c,w} = F_{tp}(X_{t,c,w}) = \frac{1}{T} \sum_{t=1}^{T} X_{t,c,w} \]  \hspace{1cm} (3)

\[ Z_{c,1,1} = \Gamma_{gap}(Y_{c,1,w}) = \frac{1}{WH} \sum_{w=1}^{W} \sum_{h=1}^{H} Y_{c,1,w} \]  \hspace{1cm} (4)

Where \( X_{t,c,w} \) denotes frame-level feature volume. It is notable that intra-clip pooling does not change the dimension of frame channel, which means this operation not generated any high-level features. Thus, we concatenate a learnable module after that to encode more distinctive clip-level features. The formulation of clip-level encoding as following:

\[ Q_{c,1,1} = G_{c}(Z_{c,1,1}) \]  \hspace{1cm} (5)

Where \( Q_{c,1,1} \) is the target clip-level feature and \( G_{c}(\cdot) \) represents the intra-clip encoding operation. To further reduce the computational cost, we adopted bottleneck structure to instead of “one-step encoding” structure in intra-clip encoding block.
Figure 2: Illustration of ICA module. The left part shows two-type of pooling operation (temporal pooling and global average pooling) to generate a temporary clip-level feature. The right part is a learnable block, which are used to generate more high-level intra-clip features. We designed the learnable block as bottleneck structure.

4 Experiments

In this section, we thoroughly evaluate the proposed intra-clip aggregation module and clip-level data augmentation technique on popular benchmark video re-id datasets. Firstly, we introduce the experimental setting and implementation details in Sec 4.1, and then conduct ablation study and present a group of visualization to support the ablation analysis in Sec 2.2. In Sec 4.3, we compare the performance of our method with state-of-the-art. At last, we perform cross-dataset evaluation to verify the generalization of the proposed method.

4.1 Experimental setting

Datasets. We conduct experiments on three commonly used video-based person re-id datasets, including ILIDS-VID, PRID 2011 and MARS. (a) The ILIDS-VID dataset is composed of 600 video sequences of 300 distinct pedestrians observed in two non-overlapping cameras views. The length of each video sequences varies from 23 to 192, with an average number of 73 frames. The train and test set are splited evenly of 150 identities. It is a challenging dataset due to random occlusion, cluttered background as well as drastically variation in illumination and blur. (b) The PRID 2011 datasets is another standard benchmark for video-based person re-id. It contains 400 image sequences for 200 pedestrians from two non-overlapping cameras. Following [3, 32], video sequences with more than 21 frames are used. Finally, the experimental dataset contains 356 image sequences for 178 identities in total. Compared with ILIDS-VID, PRID 2011 dataset, which collected in uncrowded scenes with relatively clean backgrounds and rare occlusions. (c) The MARS is the one of the largest datasets, which consists of 1,261 different IDs and around 20,000 “tracklets” (image sequence) from 6 cameras. Each IDs, which is observed by at least 2 cameras, has 13.2 tracklets on average. All tractlets are obtained by computer vision algorithm which make the dataset more challenging than datasets above. We evaluate the performance of proposed method on ILIDS-VID and MARS, and perform cross-dataset evaluation on PRID2011.
Evaluation Protocols. In our experiments, we follow the popular evaluation protocol: mean average precision (mAP) score and cumulative matching characteristic (CMC) curve. For the MARS dataset, we adopt mAP score and the CMC curve at Rank-1, Rank-5, Rank-10, Rank-20 to evaluate re-id performance. For both PRID 2011 and ILIDS-VID, Rank-1, Rank-5, Rank-10 and Rank-20 score of the CMC curve are reported.

Implementation Details. During training, we adopt Adam [13], as the optimizer to minimize the loss function with standard back-propagation, where batch size is set to 64 for MARS and 32 for ILIDS-VID since the last dataset is relatively small. There are PK clips in a batch, we set K=4 (identities), P=16 (clips of each person) and K=4, P=8 for MARS and ILIDS-VID respectively; Where all of P and K are randomly sampled. Each input sequence is augmented with synchronized transformation and re-sized the all frames of a given sequence to 224×112×3. Standard ResNet-50 [9] pretrained on ImageNet is used as the feature extractor. We set a smaller learning rate in our experiment, which initialized as 0.0004 and decreases to its $\frac{1}{200}$ every 200 epochs. The weight of ICA and classifier are initialized the same as ResNet strategy [8]. We use both triplet loss and cross-entropy loss [2] to train out network. During testing, we not use synchronized transformation techniques, except for re-size the input frames to 224×112×3. To calculate the similarity between query and gallery video, we average all clip level feature vector $f_r$ one IDs to represent the IDs feature vector $\bar{f}_p$, and then compute the metric distance with query feature vector $\bar{q}$. To prove the generality of our method, cross-data evaluation is conducted on PRID2011 by utilized the best model trained on ILIDS-VID.

4.2 Ablation Study

In this part, we conduct ablation investigation to analyze the effect of several factors upon the performance, which include the ICA module, the setting of synchronized transformation. Effectiveness of ICAN components. We set our baseline as random augmentation with only using intra-clip pooling block. The sequence length is set to T=4 for compare three settings: (1) baseline (2) ICA module (3) ICA moudle with synchronized transformation. The results are summarized in Table 1, from the up to down we add the components one by one, where intra-clip encoding block, ST respectively. intra-clip encoding means a “1x1+FC” block, and ST represents the intra-clip synchronized transformation. It can be observed that after adding a learnable block on baseline, the rank-1 accuracy is improved by 9.3% on ILIDS-VID and 1.8% on MARS. It also can be seen that “ICA+ synchronized transformation” performs better than “ICA + random augmentation” which indicates the effectiveness of the intra-clip data initialization. Analysis on synchronized transformation. Further exploration on the

| Datasets  | ILIDS-VID | MARS |
|-----------|-----------|------|
| Rank@k    | 1 5 20    | 1 5 20 | mAP  |
| baseline  | 76.7 92.0 96.7 | 84.2 94.0 97.4 | 77.8 |
| ICA       | 86.0 97.3 98.7 | 85.5 95.8 97.4 | 80.0 |
| ICA+ST    | **88.7** 98.7 **100.0** | **86.0** 95.8 97.7 | **80.8** |

Table 1: Comparison of different proposed components.

effect of different synchronized transformation techniques. The sequence length is set to T=4 and based on ICA module. We list the evaluation result on Table 2, from the up to down we synchronized each data augmentation techniques one by one, where random cropping.
random flapping and random erasing. It can be seen that, the performance of “synchronized cropping + synchronized flipping” is inferior, which implies given a slightly noise can ease the over-fitting problem. Overall, synchronized transformation gives consistent improvement over the random transformation. The improvement may come from encoding more complementary information.

To support our analysis we gives a group of visualization result in Figure 3, which illustrates the frame of ID1 to ID10 after random or synchronized transformation. We use ImageNet pre-trained ResNet50 to extract each frame feature and reduce its dimensionality to visualize [21]. We selected 2 clips from each ID and each clip contains 16 frames. We can see that the frames of intra-class clip are more assembly after synchronized transformation.

### 4.3 Comparison with State-of-the-arts

For fair comparisons the performance, we set the sequence length T=4 according to the other temporal aggregation method setting. We use our best result to compare with previous state-of-the-art results in Table 3. The first three methods [17, 19, 26] focus on jointly employing frame-level spatial features and frame-wise motion clues from a collection of optical flow to represent clip-level features. The fourth [16] contributed to use a succession of “one-step” convolution kernel to extract the spatial-temporal features simultaneously from a video-volume. The last five methods [3, 6, 15, 18, 26] concentrate on aggregating intra-clip frames over temporal dimension to represent a clip-level feature. We can see that our best result achieves consistently superior performance over the recent state-of-the-art methods.

To verify the generality of our method, we will conduct cross-dataset evaluation with our best model in Sec 4.4.

### 4.4 Cross-Dataset Testing

Due to the over-fitting trap, the model trained on one dataset would exhibit poor performance on another. To better understand the generalization performance of our method, we conducted cross-dataset experiment which where using the best model trained on ILIDS-VID and tested on PRID2011. Table 4 revealed that our results achieves consistently superior performance over other method, which demonstrates our model has certain generalization for the cross-dataset testing.

| Datasets                  | ILIDS-VID |     |     |
|---------------------------|-----------|-----|-----|
|                           | Rank@k    | 1   | 5   |
| Random Transformation     | 86.0      | 97.2| 98.7|
| +synchronized Cropping    | 87.3      | 98.3| 99.3|
| +synchronized Flipping    | 86.7      | 98.0| 98.7|
| +synchronized Erasing     | 88.7      | 98.7| 100.0|

Table 2: Exploring the synchronized transformation for ICAN on the ILIDS-VID datasets.
Datasets | ILIDS-VID | MARS | mAP
--- | --- | --- | ---
Rank@k | 1 | 5 | 20 | 1 | 5 | 20 | -
QAN | 68.0 | 86.8 | 97.4 | 73.7 | 84.9 | 91.6 | 51.7
AMOC+EpicFlow | 68.7 | 94.3 | 99.3 | 68.3 | 81.4 | 90.6 | 52.9
STSRN | 70.0 | 89.3 | 98.7 | 76.7 | 93.8 | 98.1 | -
Non-local+C3D | 81.3 | - | - | 84.3 | - | - | 77.0
ST2N | 57.7 | 81.7 | 94.1 | 79.3 | 91.1 | 96.0 | 66.8
STSRN | 70.0 | 89.3 | 98.7 | 76.7 | 93.8 | 98.1 | -
Spatiotemporal | 80.2 | - | - | 82.3 | - | - | 65.8
Att | - | - | - | 83.3 | 93.8 | 97.4 | 76.7
STIM+RRU | 84.3 | 96.8 | 99.5 | 84.4 | 93.2 | 96.3 | 72.6
Ours | 88.7 | 98.7 | 100.0 | 86.0 | 95.8 | 97.7 | 80.8

Table 3: Performance comparison with other state-of-the-art methods on ILIDS-VID and MARS datasets. “-”: no reported results.

Figure 3: Comparisons random transformation(up) with synchronized transformation(down) by visualization. ID1 to ID10 represented by different colours

5 Conclusions

In this paper, we focus on the problem of unaligned intra-clip data initialization and distinctive clip-level feature representation for video-based person re-id. To handle the above-mentioned problem, we design a novel framework for video-based person re-id, which consists of two main modules: Synchronized Transformation (ST) and Intra-clip Aggregation (ICA). The former module augments intra-clip frames with the same probability and the same operation, while the latter leverages two-level intra-clip encoding to generate more discriminative clip-level features. We conduct ablation study to analyze the effects of each module. Extensive experiments conducted on two benchmarks including MARS and ILIDS-VID demonstrate the synchronized transformation and ICA module are beneficial to intra-clip aggregation. Furthermore, we perform the cross-dataset evaluation with our best model to show the generality of our method.
Datasets | PRID 2011
--- | ---
| Rank@k | 1 | 5 | 20

| CNN-RNN | 28.0 | 57.0 | 81.0 |
| TRL | 29.5 | 59.4 | 82.2 |
| ASTPN | 30.0 | 58.0 | 85.0 |
| STSRN | 32.0 | 58.0 | 90.0 |
| **Ours** | **41.6** | **71.9** | **92.1** |

Table 4: Evaluation of the generality of the proposed method by cross-dataset testing

References

[1] Slawomir Bak and Peter Carr. One-shot metric learning for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2990–2999, 2017.

[2] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[3] Ju Dai, Pingping Zhang, Dong Wang, Huchuan Lu, and Hongyu Wang. Video person re-identification by temporal residual learning. *IEEE Transactions on Image Processing*, 28(3):1366–1377, 2019.

[4] Shengyong Ding, Liang Lin, Guangrun Wang, and Hongyang Chao. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 48(10):2993–3003, 2015.

[5] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. Person re-identification by symmetry-driven accumulation of local features. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2360–2367. IEEE, 2010.

[6] Jiyang Gao and Ram Nevatia. Revisiting temporal modeling for video-based person reid. *arXiv preprint arXiv:1805.02104*, 2018.

[7] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feed-forward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256, 2010.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[10] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[11] Martin Hirzer, Csaba Beleznai, Peter M Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *Scandinavian conference on Image analysis*, pages 91–102. Springer, 2011.
[12] Mahdi M Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1062–1071, 2018.

[13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097–1105, 2012.

[15] Shuang Li, Slawomir Bak, Peter Carr, and Xiaogang Wang. Diversity regularized spatiotemporal attention for video-based person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 369–378, 2018.

[16] Xingyu Liao, Lingxiao He, and Zhouwang Yang. Video-based person re-identification via 3d convolutional networks and non-local attention. arXiv preprint arXiv:1807.05073, 2018.

[17] Hao Liu, Zequn Jie, Karlekar Jayashree, Meibin Qi, Jianguo Jiang, Shuicheng Yan, and Jiashi Feng. Video-based person re-identification with accumulative motion context. IEEE transactions on circuits and systems for video technology, 28(10):2788–2802, 2018.

[18] Yiheng Liu, Zhenxun Yuan, Wengang Zhou, and Houqiang Li. Spatial and temporal mutual promotion for video-based person re-identification. arXiv preprint arXiv:1812.10305, 2018.

[19] Yu Liu, Junjie Yan, and Wanli Ouyang. Quality aware network for set to set recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5790–5799, 2017.

[20] Bingpeng Ma, Yu Su, and Frédéric Jurie. Local descriptors encoded by fisher vectors for person re-identification. In European Conference on Computer Vision, pages 413–422. Springer, 2012.

[21] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. Journal of machine learning research, 9(Nov):2579–2605, 2008.

[22] Niall McLaughlin, Jesus Martinez del Rincon, and Paul Miller. Recurrent convolutional network for video-based person re-identification. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1325–1334, 2016.

[23] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[24] Guanglu Song, Biao Leng, Yu Liu, Congrui Hetang, and Shaofan Cai. Region-based quality estimation network for large-scale person re-identification. In Thirty-Second AAAI Conference on Artificial Intelligence, 2018.
[25] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Pose-driven deep convolutional model for person re-identification. In Proceedings of the IEEE International Conference on Computer Vision, pages 3960–3969, 2017.

[26] Xinxing Su, Yingtian Zou, Yu Cheng, Shuangjie Xu, Mo Yu, and Pan Zhou. Spatial-temporal synergic residual learning for video person re-identification. arXiv preprint arXiv:1807.05799, 2018.

[27] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin 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.

[28] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by video ranking. In European Conference on Computer Vision, pages 688–703. Springer, 2014.

[29] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In European conference on computer vision, pages 499–515. Springer, 2016.

[30] Nicolai Wojke and Alex Bewley. Deep cosine metric learning for person re-identification. In 2018 IEEE winter conference on applications of computer vision (WACV), pages 748–756. IEEE, 2018.

[31] Shuangjie Xu, Yu Cheng, Kang Gu, Yang Yang, Shiyu Chang, and Pan Zhou. Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In Proceedings of the IEEE International Conference on Computer Vision, pages 4733–4742, 2017.

[32] Jinjie You, Ancong Wu, Xiang Li, and Wei-Shi Zheng. Top-push video-based person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1345–1353, 2016.

[33] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. Mars: A video benchmark for large-scale person re-identification. In European Conference on Computer Vision, pages 868–884. Springer, 2016.

[34] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. arXiv preprint arXiv:1708.04896, 2017.

[35] Zhen Zhou, Yan Huang, Wei Wang, Liang Wang, and Tieniu Tan. 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, pages 4747–4756, 2017.