Video-Based Convolutional Attention for Person Re-Identification

Marco Zamprogno*, Marco Passon*, Niki Martinel, Giuseppe Serra, Giuseppe Lancioni, Christian Micheloni, Carlo Tasso, and Gian Luca Foresti

Università degli Studi di Udine, Udine (UD), Italy

Abstract. In this paper we consider the problem of video-based person re-identification, which is the task of associating videos of the same person captured by different and non-overlapping cameras. We propose a Siamese framework in which video frames of the person to re-identify and of the candidate one are processed by two identical networks which produce a similarity score. We introduce an attention mechanisms to capture the relevant information both at frame level (spatial information) and at video level (temporal information given by the importance of a specific frame within the sequence). One of the novelties of our approach is given by a joint concurrent processing of both frame and video levels, providing in such a way a very simple architecture. Despite this fact, out approach achieves better performance than the state-of-the-art on the challenging iLIDS-VID dataset.

Keywords: Video-Based Person Re-Identification · Visual Attention · Convolutional Attention · LSTM · iLIDS-VID.

1 Introduction

Given an image or video of a person taken from one camera, the Re-Identification task (ReID) is the process of re-associating the person by analyzing images or videos taken from a different camera with non-overlapping field of view. Although humans can easily re-identify others by leveraging descriptors based on the person's face, height, clothing, and walking pattern, ReID is a difficult problem for a machine to solve, since it should deal with features between cameras like different lighting conditions, different point of views or person occluded by objects or other people.

Traditionally many attempts to explore the problem has been proposed for still images (*e.g.*, [1–6]), while recently some research groups have experimented approaches based on video images (*e.g.*, [7]). Using videos for Re-Identification provides several advantages over still images. The video setting is a more natural way to perform Re-Identification, as a person will normally be captured by a video camera producing a sequence of images rather than a single still image. Given the availability of sequences of images, temporal information related to a

* Indicates equal contribution.

person motion may help to disambiguate difficult cases that arise when trying to recognize a person in a different camera. Furthermore, sequences of images provide a larger number of samples of a person appearance, thus allowing a better appearance model to be built. On the other hand, this large set of information needs to be treated properly.

To address this challenge, in this paper we propose an approach to the problem of video-based person re-identification that is characterized by two main aspects. First, we propose a deep neural network architecture based on a Siamese framework [8] which evaluates the similarity of the query video to a candidate one. Second, we introduce a novel spatio-temporal attention mechanism with the aim to select relevant information from different areas of the frames of the input video, and from their evolution over time. Attention mechanisms have been largely exploited in a variety of different implementations and in many different domains of Deep Learning such as Natural Language Processing [9] and Computer Vision [10]. The intuition behind Attention in Computer Vision is to mimic the human visual process. Humans give different importance to different areas in an image as they are able to focus on 'hot' areas and neglect others [11]. This improves greatly the ability to recognize structures and patterns in otherwise flat data. Nevertheless there are relatively few attempts to use Attention in the field of Automatic Re-Identification. [12] proposes integrating a soft attention based model in a Siamese network to focus adaptively on the important local regions of an input image pair. [13] uses a spatial pyramid layer as the component attentive spatial pooling to select important regions in spatial dimension. [10] proposes a spatial attention module focused on recognizing the skeleton to identify the poses, and then a temporal module to recognize the actions.

Unlike other approaches, which use at least two separate modules to identify spatial and temporal features, we use a joint module to identify both at the same time. This allows us to define a simpler architecture which provides state-of-the-art performance on the well-known iLIDS-VID dataset.

2 Related Work

The interest for video-based Person Re-Identification has increased significantly in recent years [14]. The aim of the first works was to manually extract feature representations invariant to changes in poses, lighting conditions, and viewpoints. Using these features, they proposed distance metrics to measure the similarity between two images. In particular, one of the first studies computes the spatio-temporal stable region with foreground segmentation [15]; while [16] employs more compact spatial descriptors and color features, constructed by using the manifold geometry structure in video sequences.

With the advent of Deep Learning approaches, Convolutional Neural Networks (CNNs) have been introduced in visual recognition tasks yielding to considerable improvements in the performance [17] with respect to more classical solutions [18]. In fact, CNNs are able to extract different features from a given image, representing them as a set of output maps avoiding manual effort in fea-

ture engineering. Image-based Automatic Person Re-Identification is one of the fields in which CNNs achieved remarkable results [19–24].

However, considering that Person Re-Identification is usually done in settings that involve, for example, surveillance cameras, it is easy to argue that image-based person re-identification is no more an adequate schema to address current needs.

This led to most recent works that began exploring video-based person re-identification [25, 26, 8, 27, 13, 28–30], a setting closer to real-world applications. Videos have the advantage to contain temporal information that is potentially helpful in differentiating between persons. For example, in [8], the proposed CNN model extracts features from subsequent video frames that are fed through a recurrent final layer in order to combine frame-level features and video-level features.

Not all the parts of an image or of a video are equally important and humans place more focus only on some of them, assigning little to no importance to the rest. This attention mechanism has been adopted in a variety of applications, such as machine translation [31], action recognition [32], image recognition [33] and caption generation [34]. Recently, Attention models [32, 10] have been proposed for video and image understanding. These models assign weights to different parts of each frame, making some of them more important than others. In particular, [12] proposes integrating a spatial attention based model in a siamese network to adaptively focus on the important local parts of an input image pair.

With respect to the existing literature, [29] and [35] are the most similar to our approach. [29] uses a Recurrent Neural Network (RNN) to generate temporal attentions over frames so that the model can focus on the most discriminative ones in a video. [35] instead directly calculates the attention scores on frame-based features, using a simple architecture with two separate temporal and spatial modules. Our approach exploits a single attentive module to extract both temporal and spatial features from frames at the same time, resulting in an even simpler architecture that provides state-of-the-art performance.

3 The Proposed Approach

The proposed approach (see Fig. 1) is based on a Siamese network [8]. This schema is composed by two identical networks, or branches, in which the first is fed with the query video and the second with the candidate video to be compared. Each branch includes a sequence of modules that will be described in details in next sub-sections. The parameters of the two branches are shared. The output of the Siamese network is a value that represents the similarity of the two input video sequences in terms of the distance between their respective features vectors, which should be close to zero if they belong to the same person, close to one otherwise.

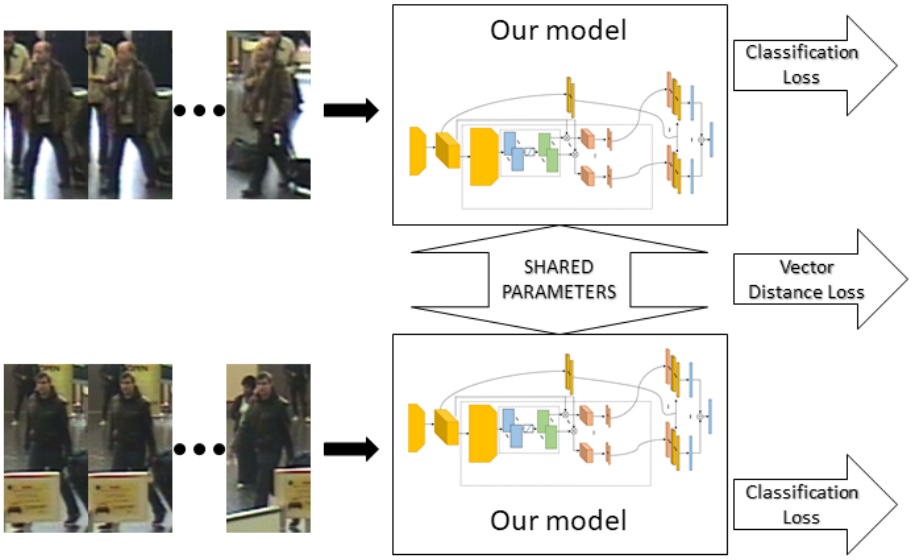

Fig. 1. Siamese network scheme. Each network receives as input a person image sequence to classify. The loss is calculated as the sum of the classification error of each network, plus the Euclidean distance between the two descriptive vectors, which should be close to zero if the two sequences belong to the same person, or close to one if they belong to different people.

3.1 Spatio-Temporal Attentive Module

The Spatio-Temporal Attentive Module is the core module of the proposed architecture. It aims to identify the portions of a frame which an human eye would normally focus on. Those areas should contain relevant spatial information, and we want to exploit them to improve the re-identification performance. Since the input frames are enhanced with the temporal information of the optical flow, both spatial and temporal features will be exploited by this network.

Inspired by [11], we propose to use a particular combination of convolutional network and LSTM, called Attentive ConvLSTM, capable of working on spatial features, in which the internal state of the network is given by the standard LSTM state equations where the matrix products between weights and inputs are replaced by convolutional operators. The ability to work with sequences is exploited to process input spatial features iteratively. The general idea of how this module works is shown in the bottom part of Fig.2.

Our aim is to exploit attentive maps to better identify relevant features of frames and provide state-of-the-art performance while using a simple network. The architecture of each branch (see Fig. 2) is based on an initial convolutional network to reduce the image size, an attentive model to generate attentive maps,

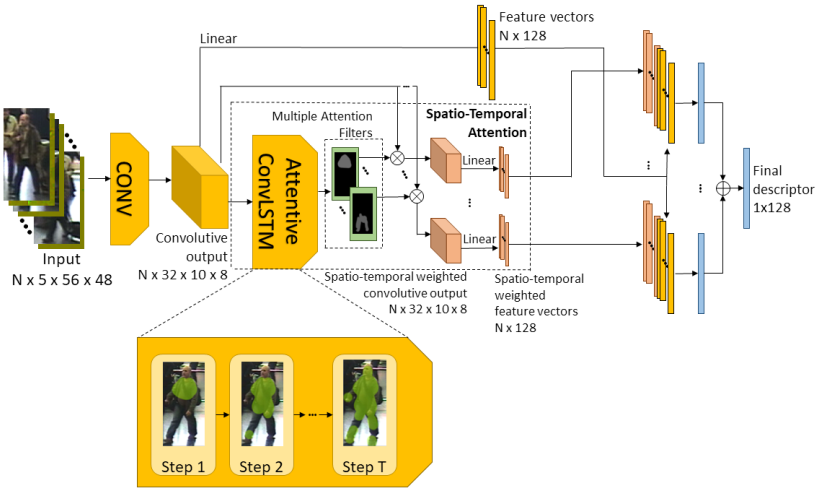

Fig. 2. Detailed Network scheme. The main blocks are the initial convolutional network, the Spatio-Temporal Attention Module, and the final part which performs averaging and normalization. The bottom part gives an idea of the multiple refining steps.

a fully connected layer to extract significant features from the original frames, and a final part where the features are combined.

More in details, the architecture consists of an ConvLSTM to recurrently processes attentive features at different locations of the frame, focusing on different regions of the tensor. A stack of features $\mathbf{X}$ is repeatedly given in input to the LSTM, which sequentially updates an internal state based on three sigmoid activators. Update is performed by two blocks: the Attentive Model, and the ConvLSTM. The Attentive Model generates an attention map using a convolutional layer that takes as input the original $\mathbf{X}$ and the previous hidden state, followed by a `tanh` activation function and another convolutional layer, and finally normalized with a `softmax` operator. The resulting output represents a normalized spatial attention map, which is then applied to the original $\mathbf{X}$ with an element-wise product, resulting in the filtered $\mathbf{X}'$. In ConvLSTM, each of the three sigmoid activators is given in input the sum of two different convolutive layers, the first taking as input $\mathbf{X}'$ and the second taking as input the previous hidden state, and a bias. The output of the first sigmoid is then multiplied element-wise with the previous $\mathbf{X}'$, the output of the second sigmoid is multiplied element-wise with the state of the LSTM memory cell, and the two resulting outputs are summed together and fed to a `tanh` activator. The result is multiplied element-wise with the output of the third sigmoid, and the resulting tensor is the new hidden state.

The Spatio-Temporal Attentive Module takes in input an image and produces in output multiple attentive maps, using an iterative refinement in T steps (based on our preliminary experiments, we set $T = 10$). We then apply those maps to

the original input and obtain multiple different filtered outputs. Ideally, each filter should focus on a different spatio-temporal feature of the frame.

3.2 Architecture Details

The starting input (see Fig. 2) consists of video sequences composed by a batch of N frames, each frame has size 56×48, with 3 channels for the YUV, plus 2 for the vertical and horizontal components of the optical flow, for a total of 5 channels.

The input is first processed through a convolutional network which consists of 3 stages, each composed by convolution, max-pooling, and nonlinear activations. Each convolution filter uses 5×5 kernels with 1×1 stride and 4×4 zero padding. This outputs a batch of size $N \times 32 \times 10 \times 8$.

At this point, the model branches in two lines: the same input is passed to the Spatio-Temporal Attentive Module previously described, and to a fully connected layer preceded by a dropout applied with $p = 0.6$ probability. The first aims to output multiple spatio-temporal-filtered feature vectors for each frame, and the second a general feature vector for each frame.

Spatio-Temporal Attention generates multiple attentive filters. Each of these filters has size 10×8, is first normalized with a sigmoid between 0 and 1, and then applied with an element-wise multiplication to the original output of the first convolutional network, obtaining multiple blocks weighted with a different filter with the same dimension of the input, $N \times 32 \times 10 \times 8$; each of these blocks focus on a specific zone of the frames. A final fully connected layer generates, for each block, a batch of spatio-temporal-weighted feature vectors of size $N \times 128$. This final layer is also preceded by a dropout with p=0.6. In our model, since we generate 3 filters, we obtain 3 spatio-temporal-weighted feature vectors.

The two branches of the network are then merged together, and the general feature vectors are concatenated with each of the spatio-temporal weighted feature vectors, resulting in 3 combined-feature vectors of size $2N \times 128$. Finally, each of these batches is averaged, normalized using L2 normalization, and lastly summed together, obtaining a final feature descriptor of size 1×128.

4 Experimental Results

Our approach has been tested and evaluated on the public iLIDS-VID benchmark [27], since it is a challenging dataset that contains many occlusions, severe illumination changes and background clutters. It is also widely used in literature and it is then easier to fair compare our results. The iLIDS-VID dataset consists of videos of 300 distinct people. For each person there are two different video sequences, captured by two non-overlapping cameras. The video sequences have a varying number of frames, with the shortest sequence having 23 frames long and the longest having 192 frames, averaging at 73 frames.

4.1 Experimental setup

To be comparable with literature, we follow the experimental setup proposed by [8]. The dataset is randomly split in two: 50% of the people form the training set and 50% the test set. During the execution of the experiments, a different train/test split is computed for every repetition and the final results are then averaged. The network is trained for 1500 epochs using Stochastic Gradient Descent algorithm. One epoch consists in showing the Siamese network an equal number of positive sequence pairs and negative pairs, sampled randomly from all the persons in the training set, alternatively.

A positive sequence pair consists of two full sequences of arbitrary length, recorded by two different cameras, showing the same person. Analogously, a negative sequence pair shows two different persons. During the training phase, the length of the sequences is set to 16, that is, 16 consecutive frames belonging to a person are randomly sampled and used during this phase. As in [27], the first camera is the probe and the second the gallery.

All the images in the dataset go through a preprocessing step where they are converted from the RGB to the YUV color space and each color channel is normalized in order to have zero mean and unit variance. The three color channels are expanded with two more channels corresponding to the horizontal and vertical component of the optical flow computed between each pair of consecutive frames using Lucas-Kanade algorithm [36]. The two optical flow channels are normalized to bring them within the $[-1, 1]$ range.

Data augmentation is applied to each sequence during the training phase in order to increase the diversity of the training sequences. Each frame in the sequence undergoes cropping and mirroring, the same transformation is applied in the same way to all the frames belonging to the same sequence.

The testing phase is performed considering a video sequence belonging to the first camera as probe and a video sequence belonging to the second camera as gallery. In this phase, we use up to to 128 frames to form a sequence. The frames are always the starting frames for the probe, and the ending frames for the gallery. If this is not possible, because a person's sequence does not have enough frames, we consider all the available frames.

All tests are performed 10 times with different seeds, each time presenting the model different people for training and testing.

4.2 Results

First we compared the results of our model when using different numbers of filters for the Spatio-Temporal Attention Module, as shown in Table 1. We found that performance increases when generating more filters, but with four or more the model saturates and the performance starts decreasing.

Second, we present experimental results with 3 filters on sequences of varying lengths between 2 and up to 128 frames, and the results are shown in Table 2. Note that if a person's sequence does not have enough frames, we still consider all the available frames and that in all cases the training has been performed

Table 1. Average results obtained using an increasing number of filters.

Average results using different number of filters				
#filters	Rank-1	Rank-5	Rank-10	Rank-20
0	60.5	84.8	93	96.9
1	59.4	85.7	93.2	97.4
2	63	**87.7**	93.9	97.3
3	**63.3**	87.4	**94**	**97.8**
4	59.6	87.2	93.9	97.7

using a fixed length sequence of 16. As one could expect, it is confirmed that the performance increases as the number of frames in sequence of frames grows, as also noted by [8]. Since the average sequence length in the dataset is 73, the performance does not increase much between 64 and 128, because most sequences are not long enough to benefit from the additional length.

Table 2. Average results with different sequence lengths (expressed in frames).

Average results with different sequence lengths				
Length	Rank-1	Rank-5	Rank-10	Rank-20
2	16.7	37.7	50.9	64.6
4	22.7	46.9	60.3	72.6
8	31.7	59.3	71.3	84.2
16	43.8	72.6	83.9	91.4
32	53.9	80.7	89	95.3
64	61	85.6	92.5	96.7
128	63.3	87.4	94	97.8

Finally, we present the comparison of our model with the state-of-the-art in Table 3. Despite beeing a simple architecture, our solution outperforms other methods proposed in the literature on 2 metrics out of 4. Note that [35] claim better results on their paper, but, in order to provide a fair comparison, we re-ran their provided code on our dataset splits. In addition, for the sake of completeness we report the performance of [37] as well, even if their testing protocol is not directly comparable with the others, as they always use all the available frames.

The simplicity of our architecture comes from the choice of making the spatial and temporal module work jointly. In fact their output is merged in order to, hopefully, get the best of the two and select only the most relevant information obtained by their combination.

5 Conclusions

We described a novel architecture which exploits a single attentive network to extract both spatial and temporal features to perform video-based person Re-

Table 3. Comparison with state-of-the-art methods.

iLIDS-VID				
Methods	Rank-1	Rank-5	Rank-10	Rank-20
Proposed Approach	**63.3**	**87.4**	94	97.8
Rao et al.[35]	62.2[1]	86.8	**94.8**	97.8
Xu et al.[38]	62	86	94	**98**
Zhang et al.[39]	60.2	85.1	-	94.2
McLaughlin et al.[8]	58	84	91	96
Zhengl et al.[40]	53	81.4	-	95.1
Yan et al.[28]	49.3	76.8	85.3	90.1
Liu et al.[37]	68[2]	86.8	95.4	97.4

Identification, providing state-of-the-art performance on the recent challenging iLIDS-VID dataset.

While the improvement obtained is not groundbreaking, the experiments confirm that employing a joint spatial and temporal attention mechanism can help pushing higher the performances in the field of person Re-Identification using only simple neural networks.

Our experiments confirms that using a longer sequence of frames brings to better performance. Analogously, one may think that using an higher number of filters will always lead to better results; however this is true up to a certain point: our experiments shows that using 3 attentive filters is better than using none, but going above this number leads to a degradation in performance.

Future work will validate the results obtained in this study performing the reported experiments on other datasets.

Acknowledgements

This project was partially supported by the FVG P.O.R. FESR 2014-2020 fund, project "Design of a Digital Assistant based on machine learning and natural language, and by the "PREscriptive Situational awareness for cooperative autoorganizing aerial sensor NETworks" project CIG68827500FB. ".

References

1. N. Martinel, C. Micheloni, and C. Piciarelli, "Distributed Signature Fusion for Person Re-Identification," in *International conference on Distributed Smart Cameras*, pp. 1–6, 2012.
2. N. Martinel, M. Dunnhofer, G. L. Foresti, and C. Micheloni, "Person Re-Identification via Unsupervised Transfer of Learned Visual Representations," in *International Conference on Distributed Smart Cameras*, pp. 1–6, 2017.

[1] These results were obtained in our tests on the code provided by the authors, and are substantially lower than claimed in the paper

[2] Results are shown for completeness, but are not directly comparable

3. G. Lisanti, N. Martinel, A. Del Bimbo, and G. L. Foresti, "Group Re-Identification via Unsupervised Transfer of Sparse Features Encoding," in *International Conference on Computer Vision*, pp. 2449–2458, 2017.
4. N. Martinel, G. L. Foresti, and C. Micheloni, "Unsupervised Hashing with Neural Trees for Image Retrieval and Person Re-Identification," in *International Conference on Distributed Smart Cameras*, 2018.
5. N. Martinel, "Accelerated low-rank sparse metric learning for person re-identification," *Pattern Recognition Letters*, vol. 112, pp. 234–240, 2018.
6. G. Lisanti, N. Martinel, C. Micheloni, A. Del Bimbo, and G. Luca Foresti, "From person to group re-identification via unsupervised transfer of sparse features," *Image and Vision Computing*, vol. 83-84, pp. 29–38, 2019.
7. L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," *CoRR*, vol. abs/1610.02984, 2016.
8. N. McLaughlin, J. Martinez del Rincon, and P. Miller, "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.
9. M. Passon, M. Comuzzo, G. Serra, and C. Tasso, "Keyphrase extraction via an attentive model," in *Italian Research Conference on Digital Libraries*, 2019.
10. S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.
11. M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "Predicting human eye fixations via an lstm-based saliency attentive model," *IEEE Transactions on Image Processing*, vol. 27, no. 10, pp. 5142–5154, 2018.
12. H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan, "End-to-end comparative attention networks for person re-identification," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3492–3506, 2017.
13. S. Xu, Y. Cheng, K. Gu, Y. Yang, S. Chang, and P. Zhou, "Jointly attentive spatial-temporal pooling networks for video-based person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4733–4742, 2017.
14. R. Vezzani, D. Baltieri, and R. Cucchiara, "People reidentification in surveillance and forensics: A survey," *ACM Comput. Surv.*, vol. 46, pp. 29:1–29:37, Dec. 2013.
15. N. Gheissari, T. B. Sebastian, and R. Hartley, "Person reidentification using spatiotemporal appearance," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2, pp. 1528–1535, 2006.
16. D. N. T. Cong, C. Achard, L. Khoudour, and L. Douadi, "Video sequences association for people re-identification across multiple non-overlapping cameras," in *International Conference on Image Analysis and Processing*, pp. 179–189, 2009.
17. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.
18. A. Rani, G. L. Foresti, and C. Micheloni, "A neural tree for classification using convex objective function," *Pattern Recognition Letters*, 2015.
19. X. Qian, Y. Fu, Y.-G. Jiang, T. Xiang, and X. Xue, "Multi-scale deep learning architectures for person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5399–5408, 2017.
20. E. Ustinova, Y. Ganin, and V. Lempitsky, "Multi-region bilinear convolutional neural networks for person re-identification," in *IEEE International Conference on Advanced Video and Signal Based Surveillance*, pp. 1–6, 2017.

21. R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang, "A siamese long short-term memory architecture for human re-identification," in *European Conference on Computer Vision*, pp. 135–153, 2016.
22. T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1249–1258, 2016.
23. L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1239–1248, 2016.
24. R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2, pp. 1735–1742, 2006.
25. Y. Li, L. Zhuo, J. Li, J. Zhang, X. Liang, and Q. Tian, "Video-based person re-identification by deep feature guided pooling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 39–46, 2017.
26. K. Liu, B. Ma, W. Zhang, and R. Huang, "A spatio-temporal appearance representation for video-based pedestrian re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3810–3818, 2015.
27. T. Wang, S. Gong, X. Zhu, and S. Wang, "Person re-identification by video ranking," in *European Conference on Computer Vision*, pp. 688–703, 2014.
28. Y. Yan, B. Ni, Z. Song, C. Ma, Y. Yan, and X. Yang, "Person re-identification via recurrent feature aggregation," in *ECCV*, pp. 701–716, 2016.
29. Z. Zhou, Y. Huang, W. Wang, L. Wang, and T. Tan, "See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification," in *CVPR*, pp. 4747–4756, 2017.
30. X. Zhu, X.-Y. Jing, X. You, X. Zhang, and T. Zhang, "Video-based person re-identification by simultaneously learning intra-video and inter-video distance metrics," *IEEE TIP*, vol. 27, no. 11, pp. 5683–5695, 2018.
31. D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.
32. S. Sharma, R. Kiros, and R. Salakhutdinov, "Action recognition using visual attention," *arXiv preprint arXiv:1511.04119*, 2015.
33. J. Ba, V. Mnih, and K. Kavukcuoglu, "Multiple object recognition with visual attention," *arXiv preprint arXiv:1412.7755*, 2014.
34. K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, pp. 2048–2057, 2015.
35. S. Rao, T. Rahman, M. Rochan, and Y. Wang, "Video-based person re-identification using spatial-temporal attention networks," 2018.
36. B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *IJCAI*, pp. 674–679, 1981.
37. Y. Liu, J. Yan, and W. Ouyang, "Quality aware network for set to set recognition," *CoRR*, vol. abs/1704.03373, 2017.
38. D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Deep metric learning for person re-identification," in *ICPR*, pp. 34–39, 2014.
39. W. Zhang, S. Hu, and K. Liu, "Learning compact appearance representation for video-based person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. abs/1702.06294, 02 2017.
40. L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian, "Mars: A video benchmark for large-scale person re-identification," in *European Conference on Computer Vision*, vol. 9910, pp. 868–884, 10 2016.