Non-local Attentive Temporal Network for Video-based Person Re-Identification

Shivansh Rao
Pennsylvania State University
State College, PA 16801, USA
shivanshrao@psu.edu

Peng Cao
University of Southern California
Los Angeles, CA 90007, USA
caopenguestc@163.com

Tanzila Rahman
University of British Columbia
Vancouver, BC V6T 1Z4, Canada
trahman8@cs.ubc.ca

Mrigank Rochan
University of Manitoba
Winnipeg, MB R3T 2N2, Canada
mrochan@cs.umanitoba.ca

Yang Wang
University of Manitoba
Winnipeg, MB R3T 2N2, Canada
ywang@cs.umanitoba.ca

Abstract

Given a video containing a person, the goal of person re-identification is to identify the same person from videos captured under different cameras. A common approach for tackling this problem is to first extract image features for all frames in the video. These frame-level features are then combined (e.g. via temporal pooling) to form a video-level feature vector. The video-level features of two input videos are then compared by calculating the distance between them. More recently, attention-based learning mechanism has been proposed for this problem. In particular, recurrent neural networks have been used to generate the attention scores of frames in a video. However, the limitation of RNN-based approach is that it is difficult for RNNs to capture long-range dependencies in videos. Inspired by the success of non-local neural networks, we propose a novel non-local temporal attention model in this paper. Our model can effectively capture long-range and global dependencies among the frames of the videos. Extensive experiments on three different benchmark datasets (i.e. iLIDS-VID, PRID-2011 and SDU-VID) show that our proposed method outperforms other state-of-the-art approaches.

1. Introduction

We consider the problem of video-based person re-identification. Given two input videos, the goal is to identify whether these two videos contain the same person. There has been lots of prior work [19, 13, 23, 22, 27] on image-based person re-identification in the literature. Given an image (probe image) with a person captured by one camera, the goal of image-based person re-identification is to match the person in a set of images (gallery images) captured by another different and non-overlapping camera. Recently, more work [18, 34, 29] has begun to focus on video-based person re-identification, since it is a more natural setting for a lot of real-world applications such as surveillance, activity analysis and tracking.

In video-based person re-identification, we are given a sequence of images rather than a static image. The key challenge is to how to exploit the temporal cues provided by the sequence. Many previous methods [18, 29, 34] usually follow a similar pipeline. First, an image-based CNN is used to extract features from each frame in the video. Then the frame-based features are aggregated to form a video-level feature vector that represents the appearance of the entire video. The distance between the video-level features of two input videos is used to indicate whether the videos contain same person or not. Ideally, the distance between videos containing the same person should be smaller, whereas the distance should be larger when the two videos contain different persons. See Fig. 1 for an illustration.

Since we now have standard techniques for extracting features from static frames, the key challenge of video-based person re-identification lies in how to assemble frame-level features into video-level features. This task is non-trivial since input videos can have variable lengths. Some previous works [34, 14] apply simple temporal pool-

978-1-5386-9294-3/18/$31.00

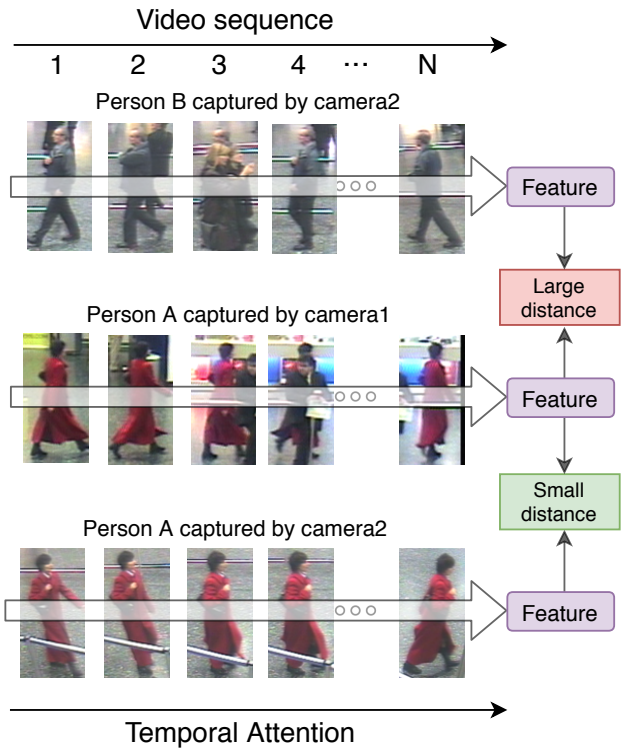

Figure 1. Illustration of video-based person re-identification problem. The problem can be formulated as learning the distance between two input videos. Our proposed method captures long-range dependencies over the entire video instead of just a local neighborhood of frames. The global dependencies could generate temporal attention to give a weight for each frame to represent its contribution to video-level features.

ing (either average or maximum pooling) over frames in get the video-level feature of a video. The limitation of simple temporal pooling is that it ignores the fact that different frames in the same video may provide different amount of information [34]. For instance, a person in one frame may be partially occluded. Intuitively, this frame may not be as informative as other frames which capture the whole body of the person. Some works [34, 29] address this issue by using recurrent neural networks (RNNs) to learn a temporal attention score for each frame in a given video sequence. The temporal attention scores give different weights to different frames in one video. Ideally, RNNs should learn to put less weights (i.e. smaller attention scores) on frames that are not informative. The limitation of the RNN approach is that it needs to perform sequential computation. As a result, it is difficult to parallelize the computation and take full advantage of GPU hardware. Moreover, a single recurrent operation could only calculate dependencies between current and latest frames. The recurrent operation has to be applied repeatedly in time. This is computationally expensive and can lead to optimization difficulties [26].

Figure 2. Some examples of the behaviour of Non-Local block in our network. The starting point of the arrows represent one frame and the ending point represents another frame of the same video sequence. These visualizations show how the model finds related clues in between the frames to support its prediction. The "blue" arrow shows similarity of first frame with the remaining frames, similarly "green", "red" and "orange" arrows are used to represent the similarity of second, third and fourth frames with the remaining frames respectively.

As a result, it is difficult for RNNs to capture long-range dependencies in a video.

In this paper, we propose a *non-local attentive temporal network* for video-based person re-identification. The novelty of our approach is that we take advantage of the recent advances in non-local neural networks [26] to compute the temporal attentions of the video frames. These temporal scores are computed in a non-local manner. Each attention score will be calculated in a way that depends on all frames in the video, not just the ones in the local neighborhood (see Fig. 2). This enables the network to effectively learn the long-range temporal dependencies among the frames in a video, and thereby improve the overall performance. Moreover, our non-local network is computationally inexpensive and can be easily paralleled to take advantage of the GPU hardware.

We demonstrate the effectiveness of our proposed method on several benchmark datasets. Our experimental results show that our method effectively captures long-ranged dependencies in videos and significantly outperforms other state-of-the-art approaches in video-based person re-identification.

2. Related Work

Person re-identification is an active area of research in computer vision. In this section, we review several lines of related work.

Image-based Person Re-identification: Given an image (probe image) including a person captured by one camera, the goal of image-based person re-identification is to match

the same person in a set of images (gallery images) captured by another different and non-overlapping camera. Existing approaches [1, 11] usually involve on two steps: (1) extracting feature vectors and (2) computing the similarity of feature vectors of two persons. To a large extent, the quality of feature vectors is crucial for the performance of person re-identification. Gray et al. [5] propose to improve viewpoint variations by using both spatial and color information. Zhao et al. [33] and Wang et al. [24] propose to use patch appearance statistics to focus on the most important parts of a person. Simonnet et al. [20] use both local and global features to capture correlated information. After extracting features from images, a distance metric is used to calculate the similarity/dis-similarity between features of two images. Ideally, the distance should be small if the two images contain the same person. Li et al. [11] propose Filter Pairing Neural Network (FPNN) to match patches across images of different views. Ahmed et al. [1] propose a deep neural network to compute distance by using cross-input neighborhood difference and patch summary structure. Shubramaniam et al. [21] uses a novel Normalized X-Corr layer to handle illuminations, occlusions and viewpoints changes.

Video-based Person Re-identification: Compared with static images, videos provide richer information for person re-identification. In addition, video-based person re-identification is closer to real-world settings. In recent years, video-based person re-identification has received lots of attention in the research community. Some earlier works [18, 7, 20] consider frame-level similarity for identifying the person. Recently, deep learning approaches are adopted to obtain more discriminative video-level features. McLaughlin et al. [18] propose a method that uses optical flows and recurrent neural networks (RNN) as well as temporal pooling layers to extract temporal information. Following [18], Xu et al. [29] propose a Spatial and Temporal Attention Pooling Network (STAPN) that computes attention scores on both spatial and temporal dimensions. The attention scores are used to get video-level features by pooling. Zhou et al. [34] also propose a network to use spatial attention and temporal attention to extract most discriminative frames and contextual information. Li et al. [10] propose a new spatiotemporal attention model to automatically discovers a diverse set of distinctive body parts.

Our work is inspired by the efficient performance of attention scores in [10, 4, 2, 31, 28] and the successful application of non-local block in video classification [26]. In this paper, we use non-local modeling to generate temporal attention scores for all the frames in a video. Compared with previous work, the novelty of our approach is that the attention scores in our method can effectively capture long-range dependencies in videos.

3. Our Approach

In our approach, we use a Siamese network architecture which takes a pair of input video sequences as its input (Fig. 3). The network architecture has two identical branches with shared model parameters. Each branch takes a video sequence as the input and produce a video-level feature vector representation that summarizes the input video. The distance of the video-level feature vectors of the two input videos is used to indicate how likely these two videos contain the same person.

Each branch of the Siamese network contains several modules. First, a frame feature extraction module is applied to extract a feature representation from each frame in the input video. Then we adopt an efficient non-local attention mechanism to assign an attention score between every pair of frames in the video. These attention scores are used to compute a weighted frame feature for each frame. For each frame, the corresponding weighted frame feature is computed based on information from all other frames in the video, so the frame feature captures non-local information of the video. Finally, a video-level feature is obtained by temporal pooling on the combination of the weighted frame features and the raw frame features.

In the following, we first describe how to extract frame-level features (Sec. 3.1). Then we introduce the non-local temporal attention module for computing the attention scores and getting the video-level feature vector (Sec. 3.2). Finally we describe the loss functions used for learning the model parameters (Sec. 3.3).

3.1. Frame Feature Extraction

Similar to [18], we use both RGB color and optical flow channels to extract the frame-level features. The color channels provide the information about the appearance of a person, while the optical flow channels provide the information about the movement of the person. Intuitively, both sources of information are useful for person re-identification. Following [18], we convert an input image (i.e. video frames) from RGB to YUV color space and normalize each color channel to have a zero mean and unit variance. To calculate both vertical and horizontal optical flow channels on each frame, we use the Lucas-Kanade algorithm [17]. Following [18], we resize each frame to have a spatial dimension of 56×40. In the end, each frame is represented as a $56 \times 40 \times 5$ tensor, where the 5 channels are composed of both 3 color channels and 2 optical flow channels.

We use the same backbone CNN architecture (Fig. 3) in [18] to extract feature-level features. It consists of three stages of convolution, max-pooling, and non-linear (tanh) activation. Each convolution filter uses 5×5 kernels with 1×1 stride and 4×4 zero padding. If the input video contains N frames, the CNN model is applied on each frame

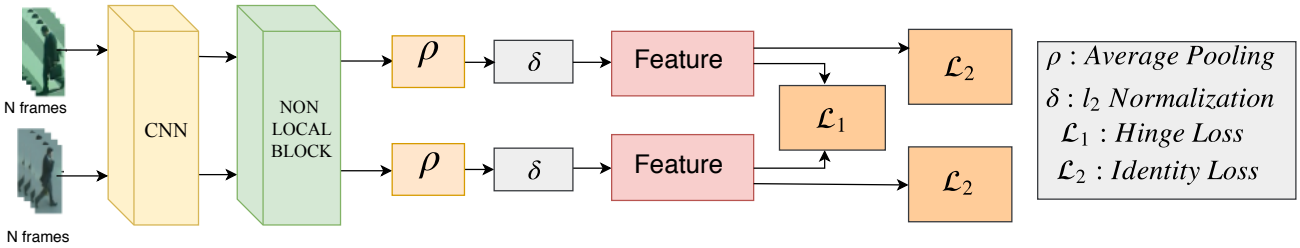

Figure 3. Illustration of the overall architecture of our proposed network. It takes a pair of input video sequence as input and passes to the Convolutional Neural Network (CNN) to extract features on each frame. The output from the CNN is fed to the Non-Local block to generate weighted feature vector for the entire video sequence. Then, average pooling and normalization are employed on the output of Non-local block to get feature vectors. Finally, the video-level feature vectors are compared to decide whether the videos contain the same person or not.

of the video to produce a $C \times H \times W$ dimensional feature $x_i \in \mathbb{R}^{C \times H \times W}$ (i= 1,2,...,N), where C is the channel dimension and $H \times W$ is the spatial dimension of the feature.

3.2. Non-local Temporal Attention

In video-based person re-identification, a key challenge is how to combine feature-level features into a video-level feature, so that the video-level features of two input videos can be compared for the re-identification. Recently, non-local neural networks [26] have been shown to be effective in capturing long-range dependencies in deep neural networks. In this paper, we use similar ideas to develop a non-local temporal attention module for person re-identification. Let x_i be a frame in a video with N frames. We use a function $f(x_i, x_j)$ (the form of $f(\cdot)$ will be defined later) to compute a scale value $y_{i,j}$ between x_i and every other frame x_j ($j \in \{1, 2, ..., N\}$). We can interpret $y_{i,j}$ as an "attention" score between these two frames. We then compute a "weighted frame feature" γ_i using the attention scores $\{y_{i,j}\}_{j=1}^{N}$ and frames $\{x_j\}_{j=1}^{N}$. Note that since γ_i is computed based on all frames in the video, γ_i implicitly contains information of the frame x_i and all the other frames in the video. To obtain the video-level feature, we simply perform temporal pooling over these weighted frame features in addition to original frame features. Since the weighted frame features already capture long-range dependencies in the video, the output (e.g. video-level feature) of the temporal pooling will implicitly capture rich long-range dependencies in the video. Figure 5 gives an illustration of the non-local temporal attention module. In the following, we provide details of the various components of this module.

Pairwise function: There are many different choices for the pairwise function f_{x_i,x_j} [26]. In our work, we adopt a version of the "dot product" with some modification for our problem. Recall that each frame is represented as a $C \times H \times W$ tensor (see Sec. 3.1), i.e. $x_i \in \mathbb{R}^{C \times H \times W}$, we define the pairwise function as follow. We apply a 1×1 convolution on x_i to reduce its channel dimension. The result is then reshaped to be a vector. We use $\theta(x_i)$ to denote this vector. The pairwise function is then defined as the dot-product between $\theta(x_i)$ and $\theta(x_j)$. In other words, the pairwise function is defined as:

$$y_{i,j} = f(x_i, x_j) = \theta(x_i)^T \theta(x_j), \text{where } \theta(x) = \text{vec}(\mathbb{C}_{1\times1}(x)) \quad (1)$$

Here (C) denotes the 1×1 convolution and $\text{vec}(\text{C})$ concatenate entries in an input tensor to form an output vector. One difference from [26] is that here we use the same function $\theta(\cdot)$ on both x_i and x_j in Eq. 1.

We then apply a softmax operation on the outputs of the pairwise function on all pairs of frames:

$$\lambda_{i,j} = \frac{\exp(y_{i,j})}{\sum_{k=1}^{N} \exp(y_{k,j})} \quad (2)$$

After the softmax operation in Eq. 2, we will have $\sum_{i=1}^{N} \lambda_{i,j} = 1$ ($\forall j$). We can interpret $\lambda_{i,j}$ as the "attention score" indicating the amount of influence of frame i on frame j.

Video-level feature: We now describe how to compute the video-level feature. First, we apply a fully connected layer on each frame x_j to produce a 128-dimensional feature vector z_j (i.e. $z_j \in \mathbb{R}^{1\times128}$). We then compute an attention weighted frame feature γ_j as follows:

$$\gamma_j = \sum_{i=1}^{N} \lambda_{i,j} z_i \quad (3)$$

Here $\gamma_j \in \mathbb{R}^{1\times128}$ is a feature vector corresponding to the j-th frame. This feature vector already incorporates the dependencies between the j-th frame and all other frames in the video. , We then concatenate $[\gamma_1, \gamma_2, ..., \gamma_N]$ with the raw frame features $[z_1, z_2, ..., z_N]$ and perform a temporal pooling [18] followed by l_2 normalization to obtain the un-

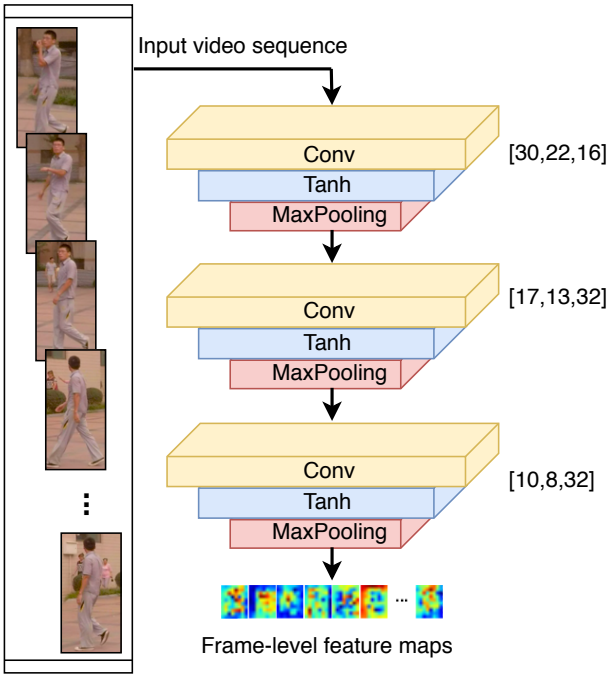

Figure 4. Our CNN architecture for extracting frame-level features. The network processes each frame (both color and optical flow channels) using a series of layers. The same CNN architecture is used in [18].

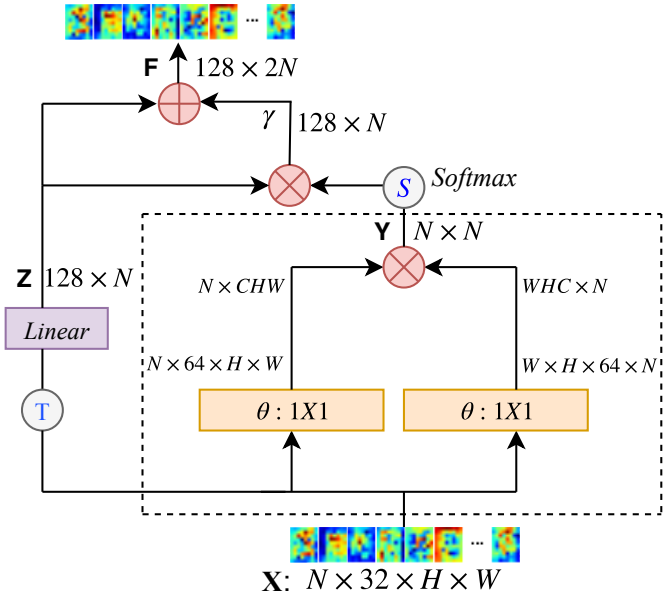

Figure 5. Illustration of the non-local temporal attention module. The dimension of the input feature maps are $N \times 32 \times H \times W$ (where 'N' represents number of frames in input video). Symbol 'T' represents transpose function. $'\otimes'$ denotes matrix multiplication and $'\oplus'$ denotes concatenation. The softmax operation is performed on each column. The yellow boxes denote 1×1 convolutions which change input features into 64 channels. Then we have used the Dot Product to calculate dependencies. The dotted box shows non-local operation and it generates $N \times N$ attention matrix.

normalized video-level feature v:

$$F = [\gamma_1, \gamma_2, ..., \gamma_N, z_1, z_2, ..., z_N] \quad (4)$$

$$v = L2\_Norm(TemporalPooling(F)) \quad (5)$$

where $TemporalPooling(\cdot)$ and $L2\_Norm(\cdot)$ denote the temporal pooling and l_2 normalization, respectively. In the end, the unnormalized video-level feature v is a 128 dimensional vector.

3.3. Model Learning

In this section, we explain the process of learning the parameters of our network. Let v_1 and v_2 be the video-level feature vectors of two input videos from the Siamese network. Similar to [18, 29], we calculate the Euclidean distance between the feature vectors and apply the squared hinge loss ($Loss_{hinge}$) as follows:

$$\mathcal{L}_{hinge} = \begin{cases} \frac{1}{2}\left\|v_1 - v_2\right\|^2, & P_1 = P_2. \\ \frac{1}{2}\lceil max(0, m - \left\|v_1 - v_2\right\|)\rceil^2, & P_1 \neq P_2. \end{cases} \quad (6)$$

Where the margin of separating two classes in $\mathcal{L}_{hinge}$ is represented by the hyper-parameter m and the identities of the persons from two input videos are represented by P_1 and P_2. The basic understanding is that if the person in both the videos is the same (i.e. P_1 = P_2) then the distance between the feature vectors should be small. Otherwise, the distance should be large, which is when the persons are different in two videos (i.e. $P_1 \neq P_2$).

Our Siamese network also has identity loss (i.e. $Loss_{id}$) added to each of its branch to predict the person's identity in a fashion similar to [18].The feature vector extracted from each of the branch in the siamese network is passed through a linear classifier to predict the person's identity. A Softmax loss is then applied over the prediction for each of the Siamese branch. The final loss is now the combination of two identity losses ($\mathcal{L}_{p1}$ and $\mathcal{L}_{p2}$) from each Siamese branch with the hinge loss as follows:

$$\mathcal{L}_{final} = \mathcal{L}_{p1} + \mathcal{L}_{p2} + \mathcal{L}_{hinge} \quad (7)$$

Stochastic gradient descent is used as the optimizer for the loss function defined in the above equation. After the training phase has been performed, all the loss functions including the identity and hinge losses are removed and then

Dataset	iLIDS-VID	PRID-2011	SDU-VID
Total no. of id.	300	749	300
No. id in multiple cameras	300	200	300
No. track-lets	600	400	600
Image resolution	64x128	64x128	64x128
No. of camera	2	2	2
Detection procedure	Hand	Hand	Hand
Evaluation Metric	CMC	CMC	CMC

Table 1. Summary of basic information of the three datasets used in our experiments.

only calculate the distance between two input video vectors for re-identification during testing.

4. Experiments

In this section, we first introduce the datasets used in the experiments (Sec. 4.1). Then we describe our experimental setup and some implementation details (Sec. 4.2). We present the experimental results and compare with other state-of-the-art in Sec. 4.3 and Sec. 4.4.

4.1. Datasets

We conduct experiments on three benchmark datasets: iLIDS-VID [25], PRID-2011 [6] and SDU-VID [15].

iLIDS-VID Dataset: This dataset consists of video sequences of 300 persons where each person is captured by a pair of non-overlapping cameras. The length of each video sequence varies from 23 to 192 frames with an average of 73 frames. This dataset is quite challenging consisting of a lot of occlusions, illumination changes, background clutters, etc.

PRID-2011 Dataset: This dataset contains video sequences of 749 persons. For the first 200 persons (or identities), there are two video sequences captured by two different cameras. The remaining persons appear in only one camera. Each sequence contains between 5 and 675 frames, with an average of 100 frames. Compared with iLIDS-VID, the PRID-2011 dataset contains fewer occlusions since the videos are captured in a relative simple environment.

SDU-VID Dataset: This dataset is similar to iLIDS-VID and PRID-2011. It contains videos captured from two non-overlapping cameras. There are 600 video sequences for 300 different pedestrian identities. Each video sequence contains 16 to 346 video frames. On average, a video contains 130 frames. This dataset is challenging for person re-identification since it contains background clutters, viewpoint variations and occlusions . This dataset contains more image frames than the iLIDS-VID and PRID-2011 datasets.

Table 1 shows the summary of these three benchmark datasets.

4.2. Setup and Implementation Details

We follow the same experimental protocol as McLaughlin et al. [18] on both datasets (iLIDS-VID and PRID-2011) in which we randomly split the dataset into two equal subsets where one subset is used for training and the other one for testing. For the SDU-VID dataset, we follow the experimental protocol of [32] with the same splitting strategy above. We have repeated all experiments 10 times for stable results. For evaluating our proposed method, we use the Cumulative Matching Characteristics (CMC) curve which is a ranking based evaluation metric. In the ideal case, the ground-truth video sequence should have the highest rank. Standard data augmentation techniques, such as cropping and mirroring, are applied to increase the amount of training data. When we train our network, to mitigate the influence of class imbalance, we consider equal number of positive and negative samples.

In the hinge loss (Eq. 6), the margin is set as m = 2. The network is trained for 3000 epochs with a batch size of one. A full epoch consists of a pair of positive and negative sample. In the non-local block, the (1x1) convolution layer within θ changes the 32 channels of input to 64 channels of output. The learning rate in the stochastic gradient descent is set to be $1e^{-4}$. For iLIDS-VID dataset, we decrease the learning rate by a factor of 10 after 1300 epochs. Whereas in PRID-2011 dataset, we decrease the learning rate two times by a factor of 10. One happens after 950 epochs and the other one happens after 1300 epochs. The value of momentum is set to be 0.9. For the SDU-VID dataset, we follow the experimental protocol of [32]. For this dataset, we also decrease the learning rate by a factor of 10 after 1300 epochs. Due to the variable-length of video sequences in these datasets, we use sub-sequences of 16 consecutive frames (N = 16) during training. If this length becomes greater than the real sequence length, then we consider the whole set of images (frames) as the sub-sequence. During testing,we consider a video sequence captured by the first camera as the probe sequence and a video sequence captured by the second camera as a gallery sequence. We use at most 128 frames in a testing video sequence. Again, if the length is greater than the real sequence, we consider the whole set of images as the video sequence. Similar strategies have been used in previous work [18].

4.3. Experimental Results

We present the results on the three benchmark datasets and compare with the other state-of-the-art methods in Table 2, Table 3 and Table 4 respectively. From the CMC rank, we can see that our method outperforms all other state-of-the-art methods by nearly 8% , 2.4% and 9% on rank-1 accuracy on iLIDS-VID, SDU-VID and PRID-2011 datasets, respectively. The comparison with [34] is particularly inter-

Method	Rank-1	Rank-5	Rank-10	Rank-20
Ours	**70**	**92**	**96**	**99**
[29]	62	86	94	98
[34]	55.2	86.5	-	97
[18]	58	84	91	96
[30]	49.3	76.8	85.3	90.1
[16]	44.3	71.7	83.7	91.7
[25]	35	57	68	78
[9]	25	45	56	66
[12]	38	63	73	82
[8]	26	48	57	69

Table 2. Comparison of our proposed approach with other state-of-the-art methods on the iLIDS-VID dataset in terms of CMC(%) at different ranks. Note that we do not include [10] in this table since it uses completely different setup and backbone network (see main text for details).

Method	Rank-1	Rank-5	Rank-10	Rank-20
Ours	**88**	**97**	**99**	**100**
[32]	85.6	97	98.3	99.6
[18]	75	86.7	-	90.8
[15]	73.3	92.7	95.3	96

Table 3. Comparison of our proposed approach with other state-of-the-art methods on the SDU-VID dataset in terms of CMC(%) at different ranks.

esting since [34] uses a similar temporal attention approach. The difference is that [34] uses RNN to generate attention scores, while our method uses the non-local block [26] to generate attention scores. The improvement of our method over [34] shows the advantage of our non-local temporal attention method. [34] additionally uses a recurrent model to generate spatial attentions. In contrast, our model only uses temporal attentions and is much simpler, yet achieves much better performance.

Some recent work in [10] has reported higher performance numbers on the iLIDS-VID and PRID-2011 datasets. However, this work has used a completely different setting from the baseline paper [18]. [10] uses ResNet-50 to extract frame-level features for each image (frame) instead of the plain CNN network (see Fig. 3.1) used by the baseline paper [18]. Moreover, [10] uses a different image size of (256,128) instead of (56,40) used in [18]. Due to these differences in experiment setup and backbone network, the accuracy numbers in [10] are not directly comparable.

4.4. Cross-Dataset Testing

In order to check the generalizability of our model, we perform cross-dataset testing following [18]. In real-world applications of person re-identification, the background of persons and the angle of cameras during testing are likely to be completely different from the training data. In order to demonstrate the generalizability of a method, a better way is to perform cross-dataset testing where the model is trained on one dataset and tested on a completely different dataset. Following [18], We use 50% of iLIDS-VID dataset for training our network, and use 50% of PRID-2011 dataset for testing.

Method	Rank-1	Rank-5	Rank-10	Rank-20
Ours	**86**	**98**	**99**	**99**
[29]	77	95	99	99
[34]	79.4	94.4	-	99.3
[18]	70	90	95	97
[30]	58.2	85.8	93.7	98.4
[16]	64.1	87.3	89.9	92
[25]	42	65	78	89
[9]	35	59	70	80
[12]	43	73	85	92
[8]	41	70	78	86

Table 4. Comparison of our proposed approach with other state-of-the-art methods on the PRID-2011 dataset in terms of CMC(%) at different ranks.Note that we do not compare with [10] because of completely different settings of backbone networks.

Method	Dataset	Rank-1	Rank-5	Rank-10	Rank-20
Ours	iLIDS-VID	**38**	**72**	**80**	**87**
[29]	iLIDS-VID	30	58	71	85
[18]	iLIDS-VID	28	57	69	81

Table 5. CMC Rank accuracy (%) using cross dataset testing (using multi-shot re-identification) on the PRID-2011 dataset. The model is trained on the iLIDS-VID dataset.

In previous methods of cross-dataset testing [18, 29], two different settings have been used for evaluation: single-shot re-identification and multi-shot re-identification. In single-shot method, only one frame of a video is used. This setting can not be applied in our work since we generate attention scores based on frame features and later combine them to produce the video-level features. So we only consider multi-shot re-identification in this paper.

In Table 5, we perform the comparison of our results with other methods using the multi-shot cross-dataset testing . Our method outperforms other methods in terms of CMC ranking accuracy by a large margin. This shows that the model learned by our method has good generalizability and can perform well on test data that are completely different from training data.

5. Conclusion

We have proposed a non-local attentive temporal network for video-based person re-identification. The main novelty of our method is a non-local temporal attention module that calculates attention scores in a global manner that considers all frames in a video. As a result, the at-

tentin scores capture long-range dependencies of all frames in a video. Our experimental results show that this global representation of video can significantly improve the performance of person re-identification.

Acknowledgment

The work was done while all authors are with the University of Manitoba. The authors acknowledge the financial support from NSERC and UMGF. Rao and Cao were supported by the Shastri research student fellowship and the MITACS globalink internship award, respectively. We thank NVIDIA for donating some of the GPUs used in this work.

References

[1] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *CVPR*, 2015.

[2] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.

[3] A. Buades, B. Coll, and J.-M. Morel. A non-local algorithm for image denoising. In *CVPR*, 2005.

[4] C. N. dos Santos, M. Tan, B. Xiang, and B. Zhou. Attentive pooling networks. *Computing Research Repository*, 2016.

[5] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, 2008.

[6] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In *SCIA*, 2011.

[7] S. Karaman and A. D. Bagdanov. Identity inference: generalizing person re-identification scenarios. In *ECCV*, 2012.

[8] S. Karanam, Y. Li, and R. J. Radke. Person re-identification with discriminatively trained viewpoint invariant dictionaries. In *ICCV*, 2015.

[9] S. Karanam, Y. Li, and R. J. Radke. Sparse re-id: Block sparsity for person re-identification. In *CVPR*, 2015.

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

[11] W. Li, R. Zhao, T. Xiao, and X. Wang. DeepReID: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014.

[12] Y. Li, Z. Wu, S. Karanam, and R. J. Radke. Multi-shot human re-identification using adaptive fisher discriminant analysis. In *BMVC*, 2015.

[13] S. Liao and S. Z. Li. Efficient PSD constrained asymmetric metric learning for person re-identification. In *ICCV*, 2015.

[14] H. Liu, Z. Jie, K. Jayashree, M. Qi, J. Jiang, and S. Yan. Video-based person re-identification with accumulative motion context. *TCSVT*, 2017.

[15] K. Liu, B. Ma, W. Zhang, and R. Huang. A spatio-temporal appearance representation for viceo-based pedestrian re-identification. In *ICCV*, 2015.

[16] K. Liu, B. Ma, W. Zhang, and R. Huang. A spatio-temporal appearance representation for video-based pedestrian re-identification. In *ICCV*, 2015.

[17] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *DARPA*, 1981.

[18] N. McLaughlin, J. M. del Rincon, and P. Miller. Recurrent convolutional neural network for video-based person re-identification. In *CVPR*, 2016.

[19] X. Qian, Y. Fu, Y.-G. Jiang, T. Xiang, and X. Xue. Multi-scale deep learning architectures for person re-identification. In *ICCV*, 2017.

[20] D. Simonnet, M. Lewandowski, S. A. Velastin, J. Orwell, and E. Turkbeyler. Re-identification of pedestrians in crowds using dynamic time warping. In *ECCV*, 2012.

[21] A. Subramaniam, M. Chatterjee, and A. Mittal. Deep neural networks with inexact matching for person re-identification. In *NIPS*, 2016.

[22] E. Ustinova, Y. Ganin, and V. Lempitsky. Multi-region bilinear convolutional neural networks for person re-identification. In *AVSS*, 2017.

[23] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang. A siamese long short-term memory architecture for human re-identification. In *ECCV*, 2016.

[24] H. Wang, S. Gong, and T. Xiang. Unsupervised learning of generative topic saliency for person re-identification. In *BMVC*, 2014.

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

[26] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. In *CVPR*, 2018.

[27] T. Xiao, H. Li, W. Ouyang, and X. Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *CVPR*, 2016.

[28] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.

[29] 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 *ICCV*, 2017.

[30] Y. Yan, B. Ni, Z. Song, C. Ma, Y. Yan, and X. Yang. Person re-identification via recurrent feature aggregation. In *ECCV*, 2016.

[31] W. Yin, H. Schutze, B. Xiang, and B. Zhou. ABCNN: Attention-based convolutional neural networks for modeling sentence pairs. *TACL*, 2016.

[32] W. Zhang, X. Yu, and X. He. Learning bidirectional temporal cues for video-based person re-identification. *TCSVT*, 2017.

[33] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *CVPR*, 2013.

[34] 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*, 2017.