Video-Based Person Re-identification via 3D Convolutional Networks and Non-local Attention

Xingyu Liao[1](✉), Lingxiao He[2], Zhouwang Yang[3], and Chi Zhang[4]

[1] School of Mathematical Sciences, University of Science and Technology of China,
Hefei, People's Republic of China
randall@mail.ustc.edu.cn

[2] University of Chinese Academy of Sciences, Beijing, People's Republic of China
lingxiao.he@nlpr.ia.ac.cn

[3] School of Data Science, University of Science and Technology of China,
Hefei, People's Republic of China
yangzw@ustc.edu.cn

[4] Megvii Inc. (Face++), Beijing, China
zhangchi@megvii.com

Abstract. Video-based person re-identification (ReID) is a challenging problem, where some video tracks of people across non-overlapping cameras are available for matching. Feature aggregation from a video track is a key step for video-based person ReID. Many existing methods tackle this problem by average/maximum temporal pooling or RNNs with attention. However, these methods cannot deal with temporal dependency and spatial misalignment problems at the same time. We are inspired by video action recognition that involves the identification of different actions from video tracks. Firstly, we use 3D convolutions on video volume, instead of using 2D convolutions across frames, to extract spatial and temporal features simultaneously. Secondly, we use a non-local block to tackle the misalignment problem and capture spatial-temporal long-range dependencies. As a result, the network can learn useful spatial-temporal information as a weighted sum of the features in all space and temporal positions in the input feature map. Experimental results on three datasets show that our framework outperforms state-of-the-art approaches by a large margin on multiple metrics.

1 Introduction

Person re-identification (ReID) aims to match people in the different places (time) using another non-overlapping camera, which has become increasingly popular in recent years due to the wide range of applications, such as public security, criminal investigation, and surveillance. Most deep learning approaches have been shown to be more effective than traditional methods. But there still remains many challenging problems because of human pose, lighting, background, occluded body region and camera viewpoints.

C. V. Jawahar et al. (Eds.): ACCV 2018, LNCS 11366, pp. 620–634, 2019.
https://doi.org/10.1007/978-3-030-20876-9_39

Video-based person ReID approaches consist of feature representation and feature aggregation. And feature aggregation attracts more attention in recent works. Although most of methods [24] (see Fig. 1(A)) propose to use average or maximum temporal pooling to aggregate features, they do not take full advantage of the temporal dependency information. To this end, RNN based methods [17] (see Fig. 1(B)) are proposed to aggregate the temporal information among video frames. However, the most discriminative frames cannot be learned by RNN based methods while treating all frames equally. Moreover, temporal attention methods [16] as shown in Fig. 1(C) are proposed to extract the discriminative frames. In conclusion, these methods mentioned above cannot tackle temporal dependency, attention and spatial misalignment simultaneously. Although there are a few methods [27] using the jointly attentive spatial-temporal scheme, it is hard to optimize the networks under severe occlusion.

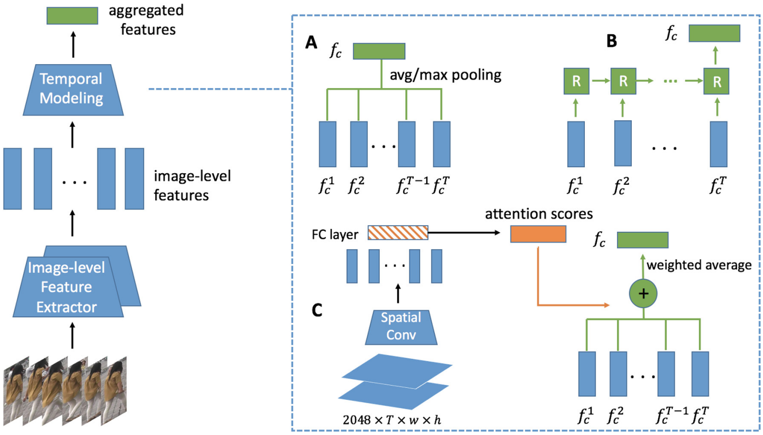

Fig. 1. Three temporal modeling methods (A: temporal pooling, B: RNN, C: temporal attention) based on an image-level feature extractor (typically a 2D CNN). For temporal pooling, average or maximum pooling is used. For RNN, hidden state is used as the aggregated feature. For attention, spatial conv + FC is shown.

In this paper, we propose a method to aggregate temporal-dependency features and tackle spatial misalignment problems using attention simultaneously as illustrated in Fig. 2. Inspired by the recent success of 3D convolutional neural networks on video action recognition [2,9], we directly use it to extract spatial-temporal features in a sequence of video frames. It can integrate feature extraction and temporal modeling into one step. In order to capture long-range dependency, we embed the non-local block [25] into the model to obtain an aggregate spatial-temporal representation.

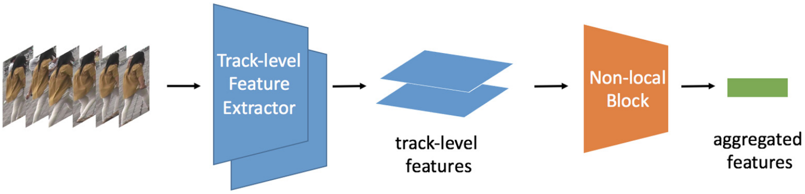

Fig. 2. The overall architecture of the proposed method. 3D convolutions are used for track-level feature extractor. Non-local blocks are embedded into aggregate spatial-temporal features.

We summarize the contributions of this work in three-folds.

1. We first propose to use 3D convolutional neural network to extract the aggregate representation of spatial-temporal features, which is capable of discovering pixel-level information and relevance among video tracks.
2. Non-local block, as a spatial-temporal attention strategy, explicitly solves the misalignment problem of deformed images. Simultaneously, the aggregative feature can be learned from video tracks by the temporal attentive scheme.
3. Spatial attention and temporal attention are incorporated into an end-to-end 3D convolution model, which achieves significant performance compared to the existing state-of-the-art approaches on three challenging video-based ReID datasets.

The rest of this paper is organized as follows. In Sect. 2, we discuss some related works. Section 3 introduces the details of the proposed approach. Experimental results on three public datasets will be given in Sect. 4. At last, we conclude this paper in Sect. 5.

2 Related Work

In this section, we first review some related works in person ReID, especially those video-based methods. Then we will discuss some related works about 3D convolution neural networks and non-local methods.

2.1 Person Re-ID

Image-based person ReID mainly focuses on feature fusion and alignment with some external information such as mask, pose, and skeleton, etc. Zhao *et al.* [29] proposed a novel Spindle Net based on human body region guided multi-stage feature decomposition and tree-structured competitive feature fusion. Song *et al.* [18] introduced the binary segmentation masks to construct synthetic RGB-Mask pairs as inputs, as well as a mask-guided contrastive attention model (MGCAM) to learn features separately from body and background regions. Suh *et al.* [20] proposed a two-stream network that consists of appearance map extraction stream and body part map extraction stream, additionally a part-aligned feature map is obtained by a bilinear mapping of the corresponding local appearance and body part descriptors. These models all actually solve the person misalignment problem.

Video-based person ReID is an extension of image-based methods. Instead of pairs of images, the learning algorithm is given pairs of video sequences. The most important part is how to fuse temporal features from video tracks. Wang *et al.* [24] aimed at selecting discriminative spatial-temporal feature representations. They firstly choosed the frames with the maximum or minimum flow energy, which is computed by optical flow fields. In order to take full use of temporal information, McLaughlin *et al.* [17] built a CNN to extract features of each frame and then used RNN to integrate the temporal information between frames, the average of RNN cell outputs are adapted to summarize the output feature. Similar to [17], Yan *et al.* [28] also used RNNs to encode video tracks into sequence features, the final hidden state is used as video representation. RNN based methods treat all frames equally, which cannot focus on more discriminative frames. Liu *et al.* [16] designed a Quality Aware Network (QAN), which is essentially an attention weighted average, to aggregate temporal features; the attention scores are generated from frame-level feature maps. In 2016, Zheng *et al.* [19] built a new dataset MARS for video-based person ReID, which becomes the standard benchmark for this task.

2.2 3D ConvNets

3D CNNs are well-suited for spatial-temporal feature learning. Ji *et al.* [9] first proposed a 3D CNN model for action recognition. Tran *et al.* [22] proposed a C3D network to be applied into various video analysis tasks. Despite 3D CNNs' ability to capture the appearance and motion information encoded in multiple adjacent frames effectively, it is difficult to be trained with more parameters. More recently, Carreira *et al.* [2] proposed the Inflated 3D (I3D) architecture which initializes the model weights by inflating the pre-trained weights from ImageNet over temporal dimension which significantly improves the performance of 3D CNNs and it is the current state-of-the-art on the Kinetics dataset [10].

2.3 Self-attention and Non-local

Non-local technique [1] is a classical digital image denoising algorithm that computes a weighted average of all pixels in an image. As attention models grow in popularity, Vaswani *et al.* [23] proposed a self-attention method for machine translation that computes the response at a position in a sequence (*e.g.,* a sentence) by attending to all positions and taking their weighted average in an embedding space. Moreover, Wang *et al.* [25] proposed a non-local architecture to bridge self-attention in machine translation to the more general class of non-local filtering operations. Inspired by these works, We embed non-local blocks into I3D model to capture long-range dependencies on space and time for video-based ReID. Our method demonstrates better performance by aggregating the discriminative spatial-temporal features.

3 The Proposed Approach

In this section, we introduce the overall system pipeline and detailed configurations of the spatial-temporal modeling methods. The whole system could be divided into two important parts: extracting spatial-temporal features from video tracks through 3D ResNet, and integrating spatial-temporal features by the non-local blocks.

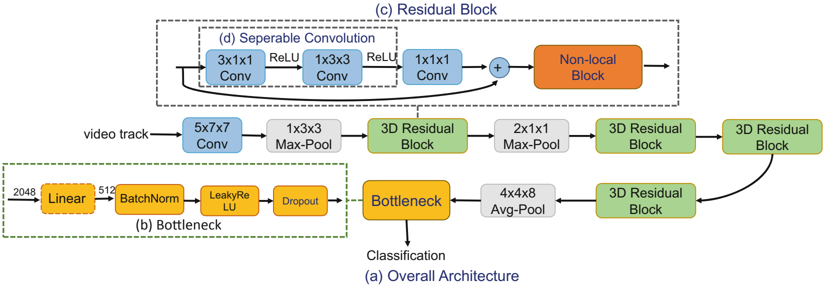

Fig. 3. Illustration of networks we propose in this paper; (a) illustrates the overall architecture that is consist of 3D convolutions, 3D pooling, 3D residual blocks, bottleneck and non-local blocks; (b) shows bottelneck; (c) illustrates residual blocks; Seperable convolutions are shown in (d).

A video tracet is first divided into consecutive non-overlap tracks $\{c_k\}$, and each track contains N frames. Supposing each track is represented as

$$c_k = \{x_t | x_t \in \mathbb{R}^{H \times W}, t = 1, \cdots, N\}, \tag{1}$$

where N is the length of c_k, and H, W are the height, width of the images respectively. As shown in Fig. 3(a), the proposed method directly accepts a whole video track as the inputs and outputs a d-dimensional feature vector f_{c_k}. At the same time, non-local blocks are embedded into 3D residual block (Fig. 3(c)) to integrate spatial and temporal features, which can effectively learn the pixel-level relevance of each frame and learn hierarchical feature representation.

Finally, average pooling followed by a bottleneck block (Fig. 3(b)) to speed up training and improve performance. A fully-connected layer is added on top to learn the identity features. A Softmax cross-entropy with label smoothing, proposed by Szegedy *et al.* [21], is built on top of the fully connected layer to supervise the training of the whole network in an end-to-end fashion. At the same time, Batch Hard triplet loss [7] is employed in the metric learning step.

During the testing, the final similarity between c_i and c_j can be measured by L2 distance or any other distance function.

In the next parts, we will explain each important component in more detail.

3.1 Temporally Separable Inflated 3D Convolution

In 2D CNNs, convolutions are applied on the 2D feature maps to compute features from the spatial dimensions only. When applied to the video-based problem, it is desirable to capture the temporal information encoded in multiple contiguous frames. The 3D convolutions are achieved by convolving 3D kernel on the cube formed by stacking multiple consecutive frames together. In other words, 3D convolutions can directly extract a whole representation for a video track, while 2D convolutions first extract a sequence of image-level features and then features are aggregated into a single vector feature. Formally, the value at position (x, y, z) on the j-th feature map in the ith layer V_{ij}^{xyz} is given by

$$V_{ij}^{xyz} = b_{ij} + \sum_{m} \sum_{p=0}^{P_i-1} \sum_{q=0}^{Q_i-1} \sum_{r=0}^{R_i-1} W_{ijm}^{pqr} V_{(i-1)m}^{(x+p)(y+q)(z+r)}, \tag{2}$$

where P_i and Q_i are the height and width of the kernel, R_i is the size of the 3D kernel along with the temporal dimension, W_{ijm}^{pqr} is the (p, q, r)th value of the kernel connected to the m-th feature map in the previous layer V_{i-1}, and b_{ij} is the bias.

We adopt 3D ResNet-50 [5] that uses 3D convolution kernels with ResNet architecture to extract spatial-temporal features. However, C3D-like 3D ConvNet is hard to optimize because of a large number of parameters. In order to address this problem, we inflate all the 2D ResNet-50 convolution filters with an additional temporal dimension. For example, a 2D $k \times k$ kernel can be inflated as a 3D $t \times k \times k$ kernel that spans t frames. We initialize all 3D kernels with 2D kernels (pre-trained on ImageNet): each of the t planes in the $t \times k \times k$ kernel is initialized by the pre-trained $k \times k$ weights, rescaled by $1/t$. According to Xie *et al.* [26] experiments, temporally separable convolution is a simple way to boost performance on variety of video understanding tasks. We replace 3D convolution with two consecutive convolution layers: one 1D convolution layer purely on the temporal axis, followed by a 2D convolution layer to learn spatial features in Residual Block as shown in Fig. 3(d). Meanwhile, we pre-train the 3D ResNet-50 on Kinetics [10] to enhance the generalization performance of the model. We replace the final classification layer with person identity outputs. The model takes T consecutive frames (*i.e.* a video track) as the input, and the layer outputs before final classification layer is used as the video track identity representation.

3.2 Non-local Attention Block

A non-local attention block is used to capture long-range dependency in space and time dealing with occlusion and misalignment. We first give a general definition of non-local operations and then provide the 3D non-local block instantiations embedded into the I3D model.

Following the non-local methods [1] and [25], the generic non-local operation in deep neural networks can be given by

$$y_i = \frac{1}{\mathcal{C}(x)} \sum_{\forall j} f(x_i, x_j) g(x_j). \tag{3}$$

Here x_i can be the position in input signal (image, sequence, video; often their features) and y_i is the position in output signal of the same size as x, whose response is to be computed by all possible input positions x_j. A pairwise function f computes a scalar between i and all j, which represents attention scores between position i in output feature and all position j in the input signal. The unary function g computes a representation in an embedded space of the input signal at the position j. At last, the response is normalized by a factor $\mathcal{C}(x)$.

Because of the fact that all positions ($\forall j$) are considered in the operation in Eq. (2), this is so-called non-local. Compared with this, a standard 1D convolutional operation sums up the weighted input in a *local* neighborhood (*e.g.*, $i-1 \leq j \leq i+1$ with kernel size 3, and recurrent operation at time i is often based only on the current and the latest time step (*e.g.*, $j = i$ or $i-1$).

There are several versions of f and g, such as gaussian, embedded gaussian, dot product, etc. According to experiments in [25], the non-local operation is not sensitive to these choices. We just choose embedded gaussian as f function that is given by

$$f(x_i, x_j) = e^{\theta(x_i)^T \phi(x_j)} \tag{4}$$

Here x_i and x_j are given in Eq. (3), $\theta(x_i) = W_\theta x_i$ and $\phi(x_j) = W_\phi x_j$ are two embeddings. We can set $\mathcal{C}(x)$ as a softmax operation, so we have a self-attention form that is given by

$$y = \sum_{\forall j} \frac{e^{\theta(x_i)^T \phi(x_j)}}{\sum_{\forall i} e^{\theta(x_i)^T \phi(x_j)}} g(x_j) \tag{5}$$

A non-local operation is very flexible, which can be easily incorporated into any existing architecture. The non-local operation can be wrapped into a non-local block that can be embedded into the earlier or later part of the deep neural network. We define a non-local block as:

$$z_i = W_z y_i + x_i \tag{6}$$

where y_i is given in Eq. (3) and "$+x_i$" means a residual connection [5]. We can plug a new non-local block into any pre-trained model, without breaking its initial behavior (*e.g.*, if W_z is initialized as zero) which can build a richer hierarchy architecture combining both global and local information.

In ResNet3D-50, we use a 3D spacetime non-local block illustrated in Fig. 4. The pairwise computation in Eq. (4) can be simply done by matrix multiplication. We will talk about detailed implementation of non-local blocks in next part.

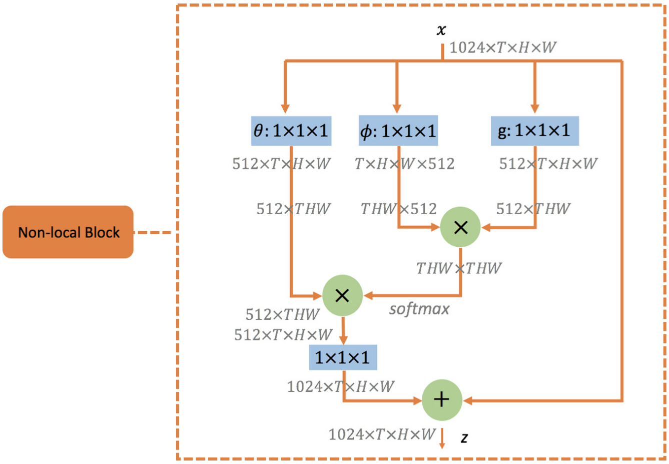

Fig. 4. The 3D spacetime non-local block. The feature maps are shown as the shape of their tensors, *e.g.*, $1024 \times T \times H \times W$ for 1024 channels (it can be different depending on networks). "⊗" denotes matrix multiplication, and "⊕" denotes element-wise sum. The softmax operation is performed on each row. The blue boxes denote $1 \times 1 \times 1$ convolutions. We show the Embedded Gaussian version, with a bottleneck of 512 channels. (Color figure online)

3.3 Loss Functions

We use triplet loss function with hard mining [7] and a Softmax cross-entropy loss function with label smoothing regularization [21].

The triplet loss function we use was originally proposed in [7], and named as Batch Hard triplet loss function. To form a batch, we randomly sample P identities and randomly sample K tracks for each identity (each track contains T frames); totally there are $P \times K$ clips in a batch. For each sample a in the batch, the hardest positive and the hardest negative samples within the batch are selected when forming the triplets for computing the loss $L_{triplet}$.

$$L_{triplet} = \overbrace{\sum_{i=1}^{P}\sum_{a=1}^{K}}^{all\ anchors} [m + \overbrace{\max_{p=1\cdots K} D(f_a^i, f_p^i)}^{hardest\ positive} \\ - \underbrace{\min_{\substack{j=1\cdots P\\ n=1\cdots K\\ j\neq i}} D(f_a^i, f_n^j)]_+}_{hardest\ negative} \tag{7}$$

The original Softmax cross-entropy loss function is given by:

$$L_{softmax} = -\frac{1}{P \times K}\sum_{i=1}^{P}\sum_{a=1}^{K} p_{i,a} \log q_{i,a} \tag{8}$$

where $p_{i,a}$ is the ground truth identity and $q_{i,a}$ is prediction of sample $\{i, a\}$. The *label-smoothing regularization* is proposed to regularize the model and make it more adaptable with:

$$L^{'}_{softmax} = -\frac{1}{P \times K} \sum_{i=1}^{P} \sum_{a=1}^{K} p_{i,a} \log((1-\epsilon)q_{i,a} + \frac{\epsilon}{N}) \tag{9}$$

where N is the number of classes. This can be considered as a mixture of the original ground-truth distribution $q_{i,a}$ and the uniform distribution $u(x) = \frac{1}{N}$.

The total loss L is the combination of these two losses.

$$L = L^{'}_{softmax} + L_{triplet} \tag{10}$$

4 Experiments

We evaluate our proposed method on three public video datasets, including iLIDS-VID [24], PRID-2011 [8] and MARS [19]. We compare our method with the state-of-the-art methods, and the experimental results demonstrate that our proposed method can enhance the performance of both feature learning and metric learning and outperforms previous methods.

4.1 Datasets

The basic information of three dataset is listed in Table 1 and some samples are displayed in Fig. 3.

Table 1. The basic information of three datasets to be used in our experiments.

Datasets	iLIDS-VID	PRID2011	MARS
#identities	300	200	1,261
#track-lets	600	400	21K
#boxes	44K	40K	1M
#distractors	0	0	3K
#cameras	2	2	6
#resolution	64 × 128	64 × 128	128 × 256
#detection	Hand	Hand	Algorithm
#evaluation	CMC	CMC	CMC & mAP

iLIDS-VID dataset consists of 600 video sequences of 300 persons. Each image sequence has a variable length ranging from 23 to 192 frames, with averaged number of 73. This dataset is challenging due to clothing similarities among people and random occlusions. **PRID-2011** dataset contains 385 persons in camera A and 749 in camera B. 200 identities appear in both cameras, constituting of 400 image sequences. The length of each image sequence varies from 5 to 675. Following [19], sequences with more 21 frames are selected, leading to 178 identities. **MARS** dataset is a newly released dataset consisting of 1,261 pedestrians captured by at least 2 cameras. The bounding boxes are generated by classic detection and tracking algorithms (DPM detector), yielding 20,715 person sequences. Among them, 3,248 sequences are of quite poor quality due to the failure of detection or tracking, significantly increasing the difficulty of person ReID.

4.2 Implementation Details and Evaluation Metrics

Training. We use ResNet3D-50 [4] as our backbone network. According to the experiments in [25], five non-local blocks are inserted to right before the last residual block of a stage. Three blocks are inserted into res_4 and two blocks are inserted into res_3, to every other residual block. Our models are pre-trained on Kinetics [10]; we also compare the models with different pre-trained weights, and the details are described in the next section.

Our implementation is based on publicly available code of PyTorch. All person ReID models in this paper are trained and tested on Linux with GTX TITAN X GPU. In training term, eight-frame input tracks are randomly cropped out from 64 consecutive frames every eight frames. The spatial size is 256 × 128 pixels, randomly cropped from a scaled videos whose size is randomly enlarged by 1/8. The model is trained on an eight-GPU machine for about 8 h, and each GPU have 16 tracks in a mini-batch (so in total with a mini-batch size of 128 tracks). In order to train hard mining triplet loss, 32 identities with 4 tracks each person are taken in a mini-batch and iterate all identities as an epoch. Bottleneck consists of fully connected layer, batch norm, leaky ReLU with $\alpha = 0.1$ and dropout with 0.5 drop ratio. The model is trained for 300 epochs in total, starting with a learning rate of 0.0003 and reducing it by exponential decay with decay rate 0.001 at 150 epochs. Adaptive Moment Estimation (Adam) is adopted with a weight decay of 0.0005 when training.

The method in [6] is adopted to initialize the weight layers introduced in the non-local blocks. A BatchNorm layer is added right after the last $1 \times 1 \times 1$ layer that represents W_z; we do not add BatchNorm to other layers in a non-local block. The scale parameter of this BatchNorm layer is initialized as zeros. This ensures that the initialize state of the entire non-local block is an identity mapping, so it can be inserted into any pre-trained networks while maintaining its initial behavior.

Testing. We follow the standard experimental protocols for testing on the datasets. For iLIDS-VID, the 600 video sequences of 300 persons are randomly split into 50% of persons for testing. For PRID2011, only 400 video sequences of the first 200 persons, who appear in both cameras are used according to experiment setup in previous methods [17] For MARS, the predefined 8,298 sequences

Table 2. Component analysis of the proposed method: rank-1, rank-5, rank-10 accuracies and mAP are reported for MARS dataset. **ResNet3D-50** is the ResNet3D-50 pre-trained on Kinectis, **ResNet3D-50 NL** is added with non-local blocks.

Methods	CMC-1	CMC-5	CMC-10	mAP
Baseline	77.9	90.0	92.5	69.0
ResNet3D-50	80.0	92.2	94.5	72.6
ResNet3D-50 NL	84.3	94.6	96.2	77.0

of 625 persons are used for training, while the 12,180 sequences of 636 persons are used for testing, including the 3,248 low quality sequences in the gallery set.

We employ Cumulated Matching Characteristics (CMC) curve and mean average precision (mAP) to evaluate the performance for all the datasets. For ease of comparison, we only report the cumulated re-identification accuracy at selected ranks.

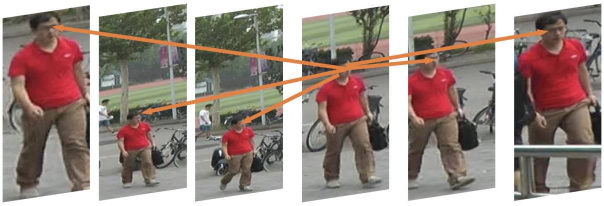

Fig. 5. Example of the behavior of a non-local block to tackle misalignment problems. The starting point of arrows represents one x_i, and the ending points represent x_j. This visualization shows how the model finds related part on different frames.

4.3 Component Analysis of the Proposed Model

In this part, we report the performance of different components in our models.

3D CNN and Non-local. Baseline method, ResNet3D-50 and ResNet3D-50 with non-local blocks on the MARS dataset are shown in Table 2. **Baseline** corresponds to ResNet-50 trained with softmax cross-entropy loss and triplet with hard mining on image-based person ReID. The representation of an image sequence is obtained by using the average temporal pooling. **ResNet3D-50** corresponds to ResNet3D-50 pre-trained on Kinetics discussed above. **ResNet3D-50 NL** corresponds to ResNet3D-50 with non-local blocks pre-trained on Kinetics. The gap between our results and baseline method is significant, and it is noted that: (1) ResNet3D increases from 77.9% to 80.0% under single query, which fully suggests ResNet3D-50 effectively aggregate the spatial-temporal features; (2) ResNet3D with non-local increase from 80.0% to 84.3% compared with

ResNet3D, which indicates that non-local blocks have the great performance on integrating spatial-temporal features and tackling misalignment problem. The results are shown in Fig. 5.

Table 3. Effect of different initialization methods: rank-1, rank-5, rank-10 accuracies and mAP are reported for MARS dataset. **ImageNet** corresponds to model pre-trained on ImageNet, **Kinetics** corresponds to model pre-trained on Kinetics and **ReID** corresponds to model pre-trained on ReID datasets.

Init Methods	CMC-1	CMC-5	CMC-10	mAP
ImageNet	78.4	91.5	93.9	69.8
ReID	79.9	92.6	94.5	71.3
Kinetics	84.3	94.6	96.2	77.0

Table 4. Comparisons of our proposed approach to the state-of-the-art on PRID2011, iLIDS-VID and MARS datasets. The rank1 accuracies are reported and for MARS we provide mAP in brackets. The best and second best results are marked by red and blue colors, respectively.

Methods	PRID2011	iLIDS-VID	MARS
AMOC+EpicFlow [15]	82.0	65.5	-
RNN [17]	40.6	58.0	-
IDE [30] + XQDA [13]	-	-	65.3 (47.3)
end AMOC+epicFlow [15]	83.7	68.7	68.3 (52.9)
Mars [19]	77.3	53.0	68.3 (49.3)
SeeForest [31]	79.4	55.2	70.6 (50.7)
QAN [16]	90.3	68.0	-
Spatialtemporal [11]	93.2	80.2	82.3 (65.8)
Ours	91.2	81.3	84.3 (77)

Different Initialization Methods. We also carry out experiments to investigate the effect of different initialization methods in Table 3. **ImageNet** and **ReID** corresponds to ResNet3D-50 with non-local block, whose weights are inflated from the 2D ResNet50 pre-trained on ImageNet or on CUHK03 [12], VIPeR [3] and DukeMTMC-reID [14] respectively. **Kinetics** corresponds to ResNet3D-50 with non-local blocks pre-trained on Kinetics. The results show that model pre-trained on Kinetics has the best performance than on other two datasets. 3D model is hard to train because of the large number of parameters and it needs more datasets to pre-train. Besides, the model pre-trained on Kinetics (a video action recognition dataset) is more suitable for video-based problem.

4.4 Comparision with State-of-the-Art Methods

Table 4 reports the performance of our approach with other state-of-the-art techniques.

Results on MARS. MARS is the most challenging dataset (it contains distractor sequences and has a substantially larger gallery set) and our methodology achieves a significant increase in mAP and rank1 accuracy. Our method improves the state-of-the-art by 2.0% compared with the previous best reported results 82.3% from Li *et al.* [11] (which use spatialtemporal attention). SeeForest [31] combines six spatial RNNs and temporal attention followed by a temporal RNN to encode the input video to achieve 70.6%. In contrast, our network architecture is straightforward to train for the video-based problem. This result suggests our ResNet3D with non-local is very effective for video-based person ReID in challenging scenarios.

Results on iLIDS-VID and PRID. The results on the iLIDS-VID and PRID2011 are obtained by fine-tuning from the pre-trained model on the MARS. Li *et al.* uses spatialtemporal attention to automatically discover a diverse set of distinctive body parts which achieves 93.2% on PRID2011 and 80.2% on iLIDS-VID. Our proposed method achieves the comparable results compared with it by 91.2% on PRID2011 and 81.3% on iLIDS-VID. 3D model cannot achieve the significant improvement because of the size of datasets. These two datasets are small video person ReID datasets, which lead to overfitting on the training set.

5 Conclusion

In this paper, we have proposed an end-to-end 3D ConvNet with non-local architecture, which integrates a spatial-temporal attention to aggregate a discriminative representation from a video track. We carefully design experiments to demonstrate the effectiveness of each component of the proposed method. In order to discover pixel-level information and relevance between each frames, we employ a 3D ConvNets. This encourages the network to extract spatial-temporal features. Then we insert non-local blocks into model to explicitly solves the misalignment problem in space and time. The proposed method with ResNet3D and non-blocks outperforms the state-of-the-art methods in many metrics.

References

1. Buades, A., Coll, B., Morel, J.M.: A non-local algorithm for image denoising. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), vol. 2, pp. 60–65, June 2005. https://doi.org/10.1109/CVPR.2005.38
2. Carreira, J., Zisserman, A.: Quo vadis, action recognition? A new model and the kinetics dataset. CoRR abs/1705.07750 (2017). http://arxiv.org/abs/1705.07750
3. Gray, D., Brennan, S., Tao, H.: Evaluating appearance models for recognition, reacquisition, and tracking. In: In IEEE International Workshop on Performance Evaluation for Tracking and Surveillance, Rio de Janeiro (2007)
4. Hara, K., Kataoka, H., Satoh, Y.: Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and imagenet? CoRR abs/1711.09577 (2017). http://arxiv.org/abs/1711.09577
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778, June 2016. https://doi.org/10.1109/CVPR.2016.90
6. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. CoRR abs/1502.01852 (2015). http://arxiv.org/abs/1502.01852
7. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person reidentification. CoRR abs/1703.07737 (2017). http://arxiv.org/abs/1703.07737
8. Hirzer, M., Beleznai, C., Roth, P.M., Bischof, H.: Person re-identification by descriptive and discriminative classification. In: Heyden, A., Kahl, F. (eds.) SCIA 2011. LNCS, vol. 6688, pp. 91–102. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-21227-7_9
9. Ji, S., Xu, W., Yang, M., Yu, K.: 3D convolutional neural networks for human action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **35**(1), 221–231 (2013). https://doi.org/10.1109/TPAMI.2012.59
10. Kay, W., et al.: The kinetics human action video dataset. CoRR abs/1705.06950 (2017). http://arxiv.org/abs/1705.06950
11. Li, S., Bak, S., Carr, P., Wang, X.: Diversity regularized spatiotemporal attention for video-based person re-identification. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018
12. Li, W., Zhao, R., Xiao, T., Wang, X.: DeepReID: deep filter pairing neural network for person re-identification. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 152–159, June 2014. https://doi.org/10.1109/CVPR.2014.27
13. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2197–2206, June 2015. https://doi.org/10.1109/CVPR.2015.7298832
14. Lin, Y., Zheng, L., Zheng, Z., Wu, Y., Yang, Y.: Improving person re-identification by attribute and identity learning. CoRR abs/1703.07220 (2017). http://arxiv.org/abs/1703.07220
15. Liu, H., et al.: Video-based person re-identification with accumulative motion context. CoRR abs/1701.00193 (2017). http://arxiv.org/abs/1701.00193
16. Liu, Y., Yan, J., Ouyang, W.: Quality aware network for set to set recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017

17. McLaughlin, N., Martinez del Rincon, J., Miller, P.: Recurrent convolutional network for video-based person re-identification. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016
18. Song, C., Huang, Y., Ouyang, W., Wang, L.: Mask-guided contrastive attention model for person re-identification. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018
19. Zheng, L., et al.: MARS: a video benchmark for large-scale person re-identification. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9910, pp. 868–884. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46466-4_52
20. Suh, Y., Wang, J., Tang, S., Mei, T., Lee, K.M.: Part-aligned bilinear representations for person re-identification. CoRR abs/1804.07094 (2018). http://arxiv.org/abs/1804.07094
21. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016
22. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3D convolutional networks. In: The IEEE International Conference on Computer Vision (ICCV), December 2015
23. Vaswani, A., et al.: Attention is all you need. In: Guyon, I., et al. (eds.) Advances in Neural Information Processing Systems, vol. 30, pp. 5998–6008. Curran Associates, Inc. (2017). http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf
24. Wang, T., Gong, S., Zhu, X., Wang, S.: Person re-identification by video ranking. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8692, pp. 688–703. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10593-2_45
25. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018
26. Xie, S., Sun, C., Huang, J., Tu, Z., Murphy, K.: Rethinking spatiotemporal feature learning for video understanding. arXiv preprint arXiv:1712.04851 (2017)
27. Xu, S., Cheng, Y., Gu, K., Yang, Y., Chang, S., Zhou, P.: Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In: The IEEE International Conference on Computer Vision (ICCV), October 2017
28. Yan, Y., Ni, B., Song, Z., Ma, C., Yan, Y., Yang, X.: Person re-identification via recurrent feature aggregation. CoRR abs/1701.06351 (2017). http://arxiv.org/abs/1701.06351
29. Zhao, H., et al.: Spindle Net: person re-identification with human body region guided feature decomposition and fusion. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017
30. Zheng, L., Zhang, H., Sun, S., Chandraker, M., Tian, Q.: Person re-identification in the wild. CoRR abs/1604.02531 (2016). http://arxiv.org/abs/1604.02531
31. Zhou, Z., Huang, Y., Wang, W., Wang, L., Tan, T.: See the forest for the trees: joint spatial and temporal recurrent neural networks for video-based person re-identification. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6776–6785, July 2017. https://doi.org/10.1109/CVPR.2017.717