Multi-Scale 3D Convolution Network for Video Based Person Re-Identification

Jianing Li, Shiliang Zhang, Tiejun Huang
School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China
{ljn-vmc, slzhang.jdl, tjhuang}@pku.edu.cn

Abstract

This paper proposes a two-stream convolution network to extract spatial and temporal cues for video based person Re-Identification (ReID). A temporal stream in this network is constructed by inserting several Multi-scale 3D (M3D) convolution layers into a 2D CNN network. The resulting M3D convolution network introduces a fraction of parameters into the 2D CNN, but gains the ability of multi-scale temporal feature learning. With this compact architecture, M3D convolution network is also more efficient and easier to optimize than existing 3D convolution networks. The temporal stream further involves Residual Attention Layers (RAL) to refine the temporal features. By jointly learning spatial-temporal attention masks in a residual manner, RAL identifies the discriminative spatial regions and temporal cues. The other stream in our network is implemented with a 2D CNN for spatial feature extraction. The spatial and temporal features from two streams are finally fused for the video based person ReID. Evaluations on three widely used benchmarks datasets, *i.e.*, *MARS*, *PRID2011*, and *iLIDS-VID* demonstrate the substantial advantages of our method over existing 3D convolution networks and state-of-art methods.

Introduction

Current researches on person Re-Identification (ReID) mainly focus on two lines of tasks depending on still images and video sequences, respectively. Recent years have witnessed the impressive progresses in image based person ReID, *e.g.*, deep visual representations have significantly boosted the ReID performance on image based ReID datasets (Li, Zhu, and Gong 2018b; Xu et al. 2018; Liu et al. 2018b; Su et al. 2016; 2015). Being able to explore plenty of spatial and temporal cues, video based person ReID has better potentials to address some challenges in image based person ReID. Fig. 1 shows several sampled frames from person tracklets. As shown in Fig. 1(a), solely relying on visual cues is hard to identify those two persons wearing visually similar clothes. However, they can be easily distinguished by gait cues. Meanwhile, video based person ReID could also leverage the latest progresses in image based person ReID. The two persons in Fig. 1(b) show similar gait cues, but can be easily distinguished by their spatial and appearance cues. It is easier to infer that, extracting and fusing spatial and temporal cues is important for video based person ReID.

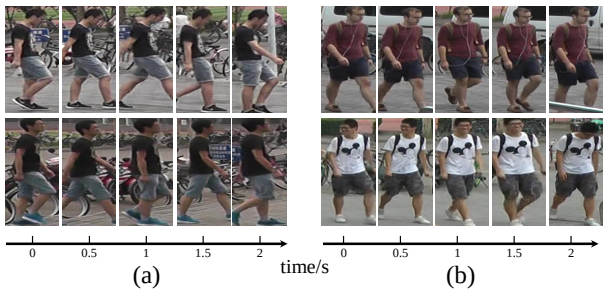

Figure 1: Illustration of video frames sampled from person tracklets. (a) shows two persons with similar appearance but different gaits; (b) shows two persons with similar gait but totally different appearance.

Existing studies on video based person ReID have significantly boosted the performance on existing datasets. Those works can be summarized into two categories, *i.e.*, 1) extracting frame-level features and generating video feature thorough pooling or weight learning (Liu et al. 2017a; Zhou et al. 2017; Li et al. 2018), and 2) extracting frame-level features then applying the Recurrent Neural Networks (RNN) to generate video features (Yan et al. 2016; McLaughlin et al. 2016). Both those two categories of methods first treat each video frame independently. The feature genearted by pooling strategy are generally not affected by the order of video frames. The RNN only builds temporal connections on high-level features, hence is not capable to capture the temporal cues on image local details. Therefore, more effective way of acquiring spatial-temporal feature should still be investigated.

Recently, 3D Convolutional Neural Network (CNN) is introduced to learn the spatial-temporal representation in other video tasks like action recognition (Carreira and Zisserman 2017; Qiu, Yao, and Mei 2017; Tran et al. 2018). Through sliding convolution kernels on both spatial and temporal dimensions, 3D CNN encodes both the visual appearance and the temporal cues across consecutive frames. Promising performances have been reported in many studies (Carreira and Zisserman 2017; Tran et al. 2015; Ji et al. 2013). Because a

single 3D convolution kernel can only cover short temporal cues, researcher usually stack several 3D kernels together to gain the stronger temporal cue learning ability. Although showing better performance, stacked 3D convolutions results in substantial growth of parameters, *e.g.*, the widely used C3D (Tran et al. 2015) network reaches the model size of 321MB with only 8 3D convolution layers, almost 3 times to the 95.7MB parameters of ResNet50 (He et al. 2016). Too many parameters not only make 3D CNNs computationally expensive, but also leads to the difficulty in model training and optimization. This makes 3D CNN not readily applicable on video based person ReID, where the training set is commonly small and person ID annotation is expensive.

This work aims to explore the rich temporal cues for person ReID through applying 3D convolution, while mitigating the shortcomings in existing 3D CNN models. A Multi-scale 3D (M3D) convolution layer is proposed as a more efficient and compact alternatives to traditional 3D CNN layer. M3D layer is implemented using several parallel temporal convolution kernels with different temporal ranges. Several M3D layers are inserted into a 2D CNN architecture. The resulting M3D convolution network (M3D CNN) introduces marginal parameters to the 2D CNN, but gains the multi-scale temporal cues modeling ability. Compared with existing 3D CNNs, M3D CNN is more compact and easier to train. To further refine the learned temporal cues by M3D convolution layer, a Residual Attention Layer (RAL) is proposed to jointly learn spatial and temporal attention masks. With RAL, more important spatial and temporal cues can be kept and the noises can be depressed, enabling M3D CNN to extract discriminative temporal feature.

We further introduce a 2D CNN to learn and extract the spatial and appearance features from video sequences. This 2D CNN and the M3D CNN compose a two-stream CNN architecture, where the extracted spatial and temporal features are fused for video based person ReID. Extensive experiments demonstrate that our method outperforms a wide range of state-of-art methods on three widely used benchmarks datasets, *i.e.*, *MARS* (Zheng et al. 2016), *PRID2011* (Hirzer et al. 2011) and *iLIDS-VID* (Wang and Zhao 2014). Moreover, we achieve a reasonable trade-off between ReID accuracy and model size. Introducing only about 4MB parameter overhead to the 2D CNN, M3D CNN boosts the mAP of 2D CNN from 0.625 to 0.699 on *MARS*. The 3D CNN model I3D (Carreira and Zisserman 2017) achieves mAP of 0.628 with 186MB parameters. Compared with I3D, M3D performs better and saves about 86MB of parameters, thus could be a better temporal feature learning model for video based person ReID.

The contribution of this work can be summarized into two aspects. 1) we propose a M3D convolution layer as a more compact and efficient alternative to 3D CNN layer. M3D layer makes multi-scale temporal feature learning with a compact neural network possible. To our best knowledge, this is the first attempt of introduce 3D convolution in person ReID. 2) we further propose the RAL to learn spatial-temporal attention masks, and use a 2D CNN to extract complementary spatial and appearance features. Those components further boost the video based person ReID performance.

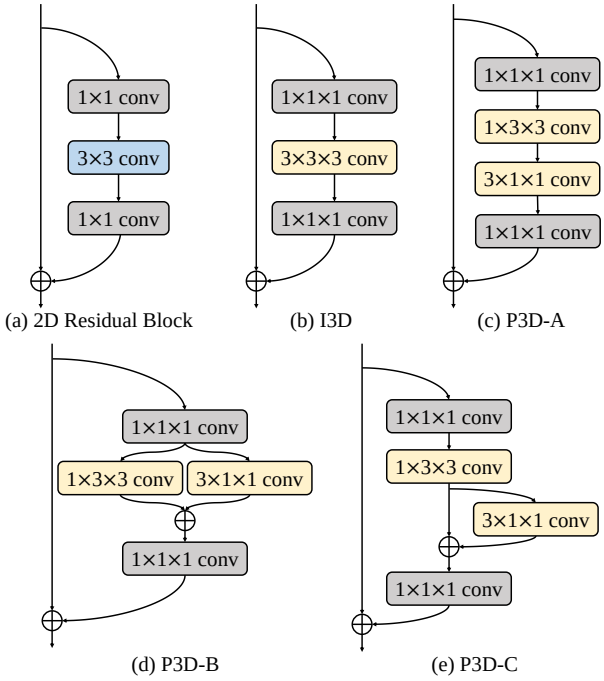

Figure 2: Some widely used convolution layer in video tasks, (a) 2D residual block; (b) I3D, which inflates 2D kernels to the 3D version; (c-e) three versions of P3D, which factorizes the 3D kernels into separate spatial and temporal ones.

Related work

Existing person ReID works can be summarized into image based person ReID and video based person ReID, respectively. Most image based person ReID works focus on two approaches: 1) learning discriminative image features (Wang and Zhao 2014; Su et al. 2017; Wei et al. 2017) and 2) learning discriminative distance metrics for feature matching (Pedagadi et al. 2013; Xiong et al. 2014). Impressive progresses have been made on image person ReID in recent years.

Many works regard video based person ReID as an extension of the image based one. For instance, 3D-SIFT (Scovanner, Ali, and Shah 2007) and HOG3D (Klaser, Marszałek, and Schmid), design hand-crafted methods to extract spatiotemporal cues, but present limited robustness when compared with deep features (Zheng et al. 2016). Some other works first extract image features from still frames, then accumulate frame features as video features. A previous work (Zheng et al. 2016) applies pooling for video feature generation. (McLaughlin et al. 2016) apply RNN to model temporal cues cross frames. (Li et al. 2018) utilize part cues and learn a weighting strategy to fuse the features extract from still frames. Unsupervised learning is also widely applied in video person ReID (Li, Zhu, and Gong 2018a; Ye, Lan, and Yuen 2018; Wu et al. 2018). Most of those works extract frame features independently and ignore the temporal cues among adjacent frames.

Some works explore both spatial and temporal cues in video tasks like action recognition through two ways. The

first one apply two-stream network to learn the spatial and temporal features, respectively, then fuse those features (Simonyan and Zisserman 2014; Feichtenhofer, Pinz, and Zisserman 2016; Feichtenhofer, Pinz, and Wildes 2017). Most of those work use still image and stacked optical flow as inputs for the two streams, respectively. The second strategy utilizes 3D CNNs to jointly explore spatiotemporal cues (Tran et al. 2015; Carreira and Zisserman 2017; Qiu, Yao, and Mei 2017; Liu et al. 2018a). Fig. 2(b-e) show 4 types of commonly used 3D CNN layer. As shown by the above works, extra temporal cues boost the performance of video tasks. However, optical flow is sensitive to the spatial misalignment between adjacent frames, which commonly exist in person ReID datasets. 3D CNNs need to stack a certain number of 3D CNN kernels to capture the long-term temporal cues. This introduces a large number of parameters and increases the difficult of 3D CNN optimization.

Our method also fuses the spatial and temporal features extracted from a two stream network. Different from previous works using stacked optical flow as input, our method directly extracts temporal feature from video sequence, hence would be more robust to the misalignment error between adjacent frames. Compared with traditional 3D CNN, our proposed M3D CNN presents better temporal cue learning ability with a more compact architecture. Those differences highlight our contribution to video based person ReID.

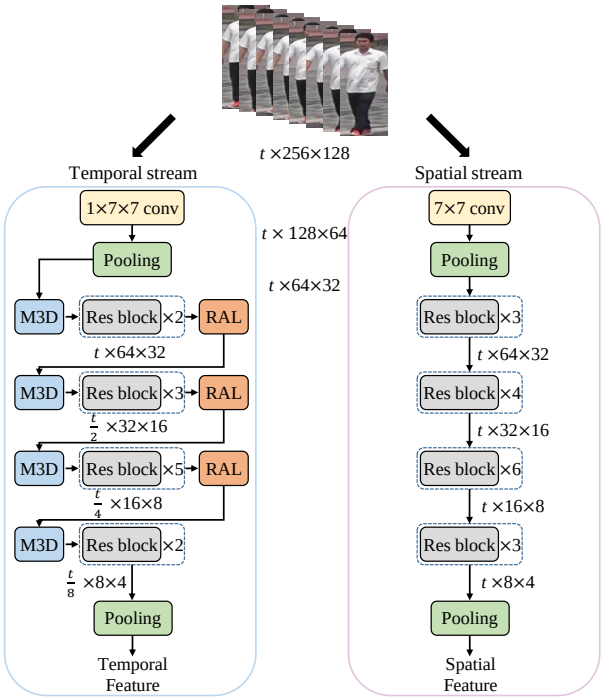

Figure 3: Illustration of Two-Stream M3D network.

Two-stream M3D Convolution Network

Problem Formulation

Person ReID aims to identify a specific person from a large scale database, which can be implemented as a retrieval task. Given a query video sequence $Q = (s^1, s^2, ...s^T)$, where T is the sequence length and s^t is the t-th frame at time t. Video based person ReID can be tackled by ranking gallery sequences based on the video representation f, and a distance metric $\mathcal{D}$ computed between Q and each gallery sequence. In the returned rank list, sequences containing the identical person with query Q are expected to appear on top of the list. Therefore, learning discriminative video representation f and designing the distance metric $\mathcal{D}$ are two critical steps for video based person ReID.

This work focuses on designing a discriminative video representation. As illustrated in Fig. 1, both the spatial and temporal cues embedded in video sequences could be important for identifying a specific person. Because the spatial and temporal cues are complementary with each other, we extract them with two modules. The video representation f_{st} can be formulated as

$$f_{st} = [f_s, f_t] \tag{1}$$

where f_s and f_t denote the spatial and temporal features, respectively, and $[,]$ denotes feature concatenation.

Existing image based person ReID works have proposed many successful methods for spatial feature extraction. As a mainstream method, 2D CNN is commonly adopted by those works. We hence refer to existing works and utilize 2D CNN to extract the sequence spatial feature f_s. Specifically, this is finished by first extracting spatial representation from each individual video frame, then aggregating frame features through average pooling, *i.e.*,

$$f_s = \frac{1}{T}\sum_{t=1}^{T} F_{2d}(s^t) \tag{2}$$

where F_{2d} refers to 2D CNN used to extract frame feature.

As discussed in the above sections, more effective ways of acquiring temporal feature should be investigated. For the temporal representation f_t, we propose a Multi-scale 3D (M3D) convolution network to learn the multi-scale temporal cues,

$$f_t = F_{M3D}(Q), \tag{3}$$

where F_{M3D} denotes the M3D network. It directly learns the temporal feature from video sequences.

The 2D CNN and M3D network compose a two stream neural network illustrated in Fig. 3. The following sections describe our design of the M3D network, which is implemented as the temporal stream in Fig. 3.

M3D Convolution Network

As illustrated in Fig. 3, the main differences between M3D network and 2D CNN are the presences of M3D and RAL. Those two layers enables M3D network to process video sequences and learn extra temporal cues. Before introducing M3D and RAL, we first briefly review 3D convolution.

3D Convolution A video clip can be represented as a 4D tensor with the size of $C \times T \times H \times W$, where C, T, H, and W denote the number of color channels, temporal length, height and width of each frame, respectively. A 3D convolution

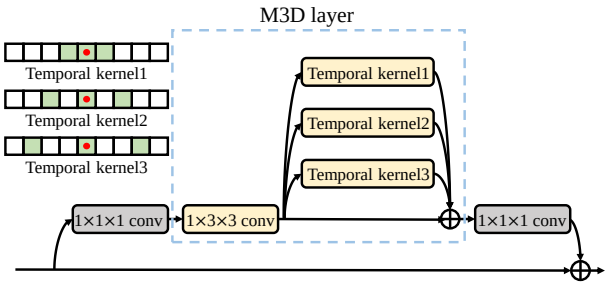

Figure 4: Illustration of M3D layer build in residual block with three temporal kernels, *i.e.*, $n = 3$.

kernel can be formulated as a 3D tensor with size of $t \times h \times w$ (the channel dimension is omitted for simplicity), where t is the temporal depth of kernel, while h and w are the spatial sizes. The 3D convolution encodes the spatial-temporal cues through sliding along both the spatial and temporal dimensions of the video clip.

3D convolution kernel only captures the short-term temporal cues, *e.g.*, the 3D kernels in Fig. 2 (b-e) capture the temporal relations across 3 frames. To model longer-term temporal cues, multiple 3D convolution kernels have to be concatenated as a deep network. A deep 3D CNN involves a large amount of parameters. Moreover, 3D CNNs can not leverage the 2D images in ImageNet (Deng et al. 2009) for model pre-training, making it further difficult to be optimized. The following parts show how our M3D layer mitigates those shortcomings in 3D convolution.

Multi-scale 3D Convolution Shortcomings of 3D CNN motivate us to design a compact convolution kernel that captures longer-term temporal cues. Inspired by dilated convolution (Yu and Koltun 2015), we propose to capture temporal cues through parallel dilated convolutions on the temporal dimension.

A M3D layer contains a spatial convolution kernel and n parallel temporal kernels with different temporal ranges. Given an input feature map $x \in R^{C\times T\times H\times W}$, we define the output of M3D layer as:

$$y = \mathcal{S}(x) + \sum_{i=1}^{n} \mathcal{T}^{(i)}(\mathcal{S}(x)) \tag{4}$$

where $\mathcal{S}$ is the spatial convolution and $\mathcal{T}^{(i)}$ is the temporal convolution with dilation rate i. The computation of S follows the ones in 2D convolution. We define the computation of $\mathcal{T}^{(i)}$ as

$$y = \mathcal{T}^{(i)}(x), \quad y_{t,h,w} = \sum_{a=-1}^{1} x_{t+a\times i,h,w} \times \mathbf{W}^{(i)}, \tag{5}$$

where $\mathbf{W}^{(i)}$ denotes the i-th temporal kernel.

Fig. 4 illustrates the detailed structure of M3D layer with $n = 3$ in residual block. As shown in Fig. 4, n controls the receptive field size in time dimension. If we set $n = 1$, the M3D layer equals to P3D-C (Qiu, Yao, and Mei 2017), *i.e.*, factorizing the 3D convolutional kernels into a spatial kernel and a temporal kernel. To limit the receptive field size no larger than the temporal dimension of input signal, given an input feature map with temporal dimension of T, we compute the number of temporal kernels n as,

$$n = \lfloor \frac{T-1}{2} \rfloor \tag{6}$$

where $\lfloor\rfloor$ means rounded down operation.

As shown in Fig. 4, with $n = 3$, M3D layer has a larger temporal receptive field than 3D convolution, *e.g.*, covering 7 time dimensions. Another advantage of M3D layer is the learning of rich long and short temporal cues through introducing multiple temporal kernels. Moreover, any 2D CNN layer can become a M3D layer by inserting temporal kernels through a residual connection as shown in Fig. 4. This structure allows M3D layer can be initialized with well-trained 2D CNN layers. For example, M3D layer can be initialized by setting weights of temporal kernels to 0, which equals to a 2D CNN layer. Initialized on a good 2D CNN model, M3D CNN would be easier to be optimized.

Residual Attention Layer In a long video sequence, different frames may present different visual qualities. Temporal cues extracted on some consecutive frames could be more important or robust than the others. Therefore, it is not reasonable to treat different spatial and temporal cues equally. We hence propose attention selection mechanisms to refine spatial and temporal cues learned by M3D layer.

We propose a Residual Attention Layer (RAL) to learn the spatial-temporal attention masks. Given an input tensor $x \in R^{C\times T\times H\times W}$, the RAL computes a saliency attention mask $M \in R^{C\times T\times H\times W}$ of the same size as x. Traditional attention masks are commonly multiplied on feature map to emphasize important local regions. As shown in (Li, Zhu, and Gong 2018b), solely emphasizing local regions and discarding the global cues may degrade the ReID performance (Li, Zhu, and Gong 2018b). To chase a more effective attention mechanism, we design the attention model with a residual manner, *i.e.*,

$$y = \frac{1}{2}x + M \cdot x, \tag{7}$$

where x and y donate the input and output 4D signals, respectively. M is the 4D attention mask which has been normalized to $(0, 1)$ by sigmoid function. As shown in Eq. 7, RAL is implemented as a residual convolution layer, where the initial input x is kept, meanwhile the meaningful cues in x are emphasized by the learned mask M.

Directly learning M can be expensive because it may contain a large number of parameters. As shown in Fig. 5, we learn M by factorising it into three low-dimensional attention masks to decrease the number of parameters, *i.e.*,

$$M = Sigmoid(S_m \times C_m \times T_m) \tag{8}$$

where $S_m \in R^{1\times 1\times H\times W}$, $C_m \in R^{C\times 1\times 1\times 1}$ and $T_m \in R^{1\times T\times 1\times 1}$ represent the spatial, channel and temporal attention masks, respectively. To learns the three masks, RAL introduces three branches, whose outputs are finally multiplied as M.

Spatial Attention Mask Learning: Spacial attention branch consists of a global temporal pooling layer and

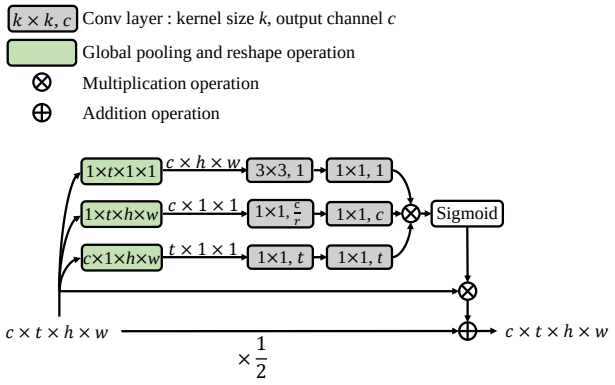

Figure 5: Illustration of Residual Attention Layer (RAL). RAL consists of three branches to apply spacial, temporal, and channel attentions. The ReLU and Batch Normalisation (BN) layer are applied after each convolution layer.

two convolution layers to compute S_m. Giving an input $x \in R^{C\times T\times H\times W}$, we define global temporal pooling as

$$x_s = \frac{1}{T}\sum_{t=1}^{T} x_{1:C,t,1:H,1:W}. \tag{9}$$

The global temporal pooling layer is designed to aggregate the information across different time dimensions. It also decreases the number of subsequent convolution parameters. We hence compute the spatial attention mask based on x_s.

A previous work (Li, Zhu, and Gong 2018b) directly averages feature maps across different channels as the spatial attention map. To model the difference across channels, we utilize a convolution layer $conv_1^s$ to generate a one-channel attention map. An 1×1 convolution layer is further introduced to learn a scale parameter for further fusion. The computation of S_m can be denoted as

$$S_m = conv_2^s(ReLU(conv_1^s(x_s))). \tag{10}$$

Channel Attention Mask Learning: The channel attention branch also contains 1 pooling layer and two 1×1 convolution layers. The first global pooling operation is imposed on spatial and temporal dimensions to aggregate the spatial and temporal cues, *i.e.*,

$$x_c = \frac{1}{T\times H\times W}\sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{w=1}^{W} x_{1:C,t,h,w}. \tag{11}$$

We then follow the Squeeze-and-Excitation (SE) (Hu, Shen, and Sun 2017) and design the channel branch with bottleneck manner:

$$C_m = conv_2^c(ReLU(conv_1^c(x_c))), \tag{12}$$

where the output channels of $conv_1^c$ is set as $\frac{c}{r}$, r represents the bottleneck reduction rate. And the output channels of $conv_2^c$ is set as c. The SE design reduces the parameters of two convolution layers from $(c^2 + c^2)$ to $\frac{1}{r}(c^2 + c^2)$, where the r is set as 16 in our experiment.

Temporal Attention Mask Learning: The temporal attention branch has the same architecture as the channel attention branch. It first aggregates the spacial and channel dimensions through global pooling. Then the attention mask can be obtained through two convolution layers.

The outputs from three branches are combined as the final attention mask M, which is further normalized to $[0, 1]$ through sigmoid function. Through initializing all convolution layers as zero, we can get $M = \frac{1}{2}$. Finally M is imposed to the input feature map with a residual manner as Eq. 7.

With the designed M3D layer and RAL, we build M3D convolution network based on ResNet50 as illustrated in Fig 3. More details of our network structure can be found in the following sections.

Experiment

Dataset

We use three video ReID datasets as our evaluation protocols, including *PRID*-2011 (Hirzer et al. 2011), *iLIDS-VID* (Wang and Zhao 2014) and *MARS* (Zheng et al. 2016).

PRID-2011 consists of 400 sequences of 200 pedestrians from two cameras. Each sequence has a length between 5 and 675 frames. Following the implementation in previous works (Wang and Zhao 2014; Li et al. 2018), we randomly split this dataset into train/test identities. This procedure is repeated 10 times for computing averaged accuracies.

iLIDS-VID consists of 600 sequences of 300 pedestrians from two non-overlapping cameras. Each sequence has a variable length between 23 and 192 frames. We also follow the implementation in (Wang and Zhao 2014; Li et al. 2018), randomly split this dataset into train/test identities 10 times.

MARS consists of 1261 pedestrians and 20,715 sequences under 6 cameras. Each pedestrian is captured by at least 2 cameras. This dataset provides fixed training and testing sets, which contain 630 and 631 pedestrians, respectively.

Implementation Details

We employ ResNet50 (He et al. 2016) as a simple 2D CNN baseline. All the 3D CNN models tested in this paper are build based on ResNet50 by replacing the 2D convolution layers with corresponding 3D convolution layers. M3D CNN is constructed based on ResNet50 by replacing portions of its 2D convolution layers with M3D layer. 3 RAL are further inserted to learn attention masks as illustrated in Fig. 3. We totally replace 4 residual blocks at the beginning of each stage in ResNet50.

Our model is trained and fine-tuned with PyTorch. Stochastic Gradient Descent (SGD) is used to optimize our model. Input images are resized to 256× 128. The mean value is subtracted from each (B, G, and R) channel. For 2D CNN baseline training, each batch contains 128 images. The initial learning rate is set as 0.001, and is reduced ten times after 10 epoches. The training is finished after 20 epoches. For 3D model training, we sample T adjacent frames from each video sequences as network input in each training epoch, and totally train the 3D models for 400 epoches. For length $T = 8$, we set the batch size as 24. The batch size is set as

Table 1: Comparison between M3D convolution layer and convolution layers on *MARS*.

Method	Input Frames	mAP	r1	Speed	Params
2D CNN	1	62.54	76.43	796 frame/s	95.7MB
I3D	8	62.84	76.62	81.0 clip/s	186.3MB
	16	61.58	75.11	38.7 clip/s	
P3D-A	8	60.69	75.08	90.1 clip/s	110.9MB
	16	60.52	75.69	46.9 clip/s	
P3D-B	8	67.03	79.06	93.9 clip/s	110.9MB
	16	65.07	77.63	48.7 clip/s	
P3D-C	8	67.06	79.08	87.6 clip/s	110.9MB
	16	65.17	79.44	45.4 clip/s	
M3D	8	**69.90**	**81.01**	**98.3 clip/s**	**99.9MB**
	16	66.23	80.13	49.1 clip/s	

12 for $T = 16$. The initial learning rate is set as 0.01, and is reduced ten times after 300 epoches.

During testing, we use 2D CNN to extract feature from each still frame, then fuse frame features into spatial feature through average pooling. For 3D CNN models, we sample T adjacent frames from original sequences as input. For a sequence of length L, we can get $\lfloor \frac{L}{T} \rfloor$ sampled inputs and corresponding features, respectively, where $\lfloor \rfloor$ refers to rounded down operation. The sequence-level feature is finally acquired by averaging those features. All of our experiments are implemented with GTX TITAN X GPU, Intel i7 CPU, and 128GB memory.

Evaluation on Individual Components

1) M3D Convolution Layer To verify the effectiveness of M3D convolution layer, we build a M3D CNN based on ResNet50, following the structure of temporal stream illustrated in Fig 3. To show the performance gains of M3D layers, RAL is not inserted in this experiment. We also compare several widely used temporal feature extraction methods.

I3D (Carreira and Zisserman 2017) inflates 2D kernels into 3D versions to acquire the temporal cues learning ability. Fig 2 (a-b) show the inflation process. 2D kernels are typically square, therefore they are inflated cubically, *e.g.*, $N \times N$ to $N \times N \times N$, which introduces a large mount of parameters to I3D.

P3D (Qiu, Yao, and Mei 2017) factorizes the 3D kernels into separate spatial and temporal ones, *e.g.*, factorize a $N \times N \times N$ kernel to a $1 \times N \times N$ spatial kernel and a $N \times 1 \times 1$ temporal kernel to reduce the amount of parameters. Fig. 2(c-e) shows 3 ways of factorizations, which are named as P3D-A, P3D-B and P3D-C, respectively. The factorization substantially decreases the parameters in 3D CNNs, while still need to stack many temporal kernels to capture long-term temporal cues.

We apply ResNet50 as the 2D CNN baseline. All of the 3D CNNs are implemented based on ResNet50, by replace 2D convolution layers with corresponding 3D versions. The comparison results are shown in Table 1.

I3D shows promising performance on video action recognition tasks (Carreira and Zisserman 2017). However, it achieves similar performance with 2D CNN in Table 1. The reason might be because I3D model has too many parameters, making it hard to train on the relatively small person ReID training sets. The P3D-A shows poor performance compared with 2D CNN. This could be caused by the serial connection between spatial kernel and temporal kernel, which increases the nonlinearity of the CNN model and makes it hard to be optimized. The P3D-B and P3D-C connect the spatial and temporal kernels through parallel or residual connections. They get substantial performance improvements over the 2D CNN. This shows the advantages of 3D CNN over 2D CNN in person ReID.

Table 2: The performance of M3D CNN on three datasets by inserting RAL and fusing of spatial and temporal features.

Dataset	*MARS*		*PRID*	*iLIDS-VID*
Method	mAP	r1	r1	r1
2D baseline	62.54	76.43	82.02	49.33
M3D	69.90	81.01	87.64	70.00
M3D+RAL(s)	71.04	82.19	89.89	71.33
M3D+RAL(t)	70.66	81.81	88.76	71.33
M3D+RAL(c)	71.30	82.13	89.89	72.00
M3D+RAL	71.76	82.79	91.03	72.67
Two-stream M3D	**74.06**	**84.39**	**94.40**	**74.00**

The experimental results also show that, our M3D CNN constantly outperforms 2D CNNs and other 3D CNNs. It outperforms 2D CNN and P3D-C by about 7.4% and 2.9% in mAP, respectively. Meanwhile, M3D CNN is also more compact than the compared 3D CNNs. M3D CNN contains 4 M3D layers, which bring only 4.2MB parameter overhead into 2D CNN. It is more compact than I3D (90.6MB) and P3D (15.2MB). With 8-frames clip as model input, M3D CNN achieves the speed of 98.3 clips/s (786.4 frames/s), which is also the fastest among the compared 3D CNNs. We further tested replacing all the 2D convolution layers in ResNet50 as M3D layers, but don't get further performance improvement. This implies that a small number of M3D layers already captures the long-term temporal cues in video sequences. We hence could conclude that, M3D convolution layer presents promising ability in learning multi-scale temporal cues.

It is also interesting to observe that, 3D CNNs trained with 8-frame clips outperform the ones trained with 16-frame clips. The reason could be because 16-frame clips take more memory and result in smaller batch size for training. Based on the this observation, we adopt 8-frame clips for training in the following experiments.

2) Residual Attention Layer This part further verifies the effectiveness of Residual Attention Layer (RAL), which include spatial, temporal, and channel branches. Experimental results are shown in Table 2. In the table, "2D baseline" denotes the performance of 2D ResNet50. "M3D" denotes M3D CNN without RAL. "RAL(s)", "RAL(t)", and "RAL(c)" donate attention layers only with spatial, temporal, and channel branches, respectively. "RAL" donates the complete attention layer containing 3 branches.

Table 3: Comparison with recent works on *MARS*.

Method	mAP	r1	r5	r20
BoW+kissme (Zheng et al. 2016)	15.50	30.60	46.20	59.20
LOMO+XQ (Zheng et al. 2016)	16.40	30.70	46.60	60.90
IDE+XQDA (Zheng et al. 2016)	47.60	65.30	82.00	89.00
LCAR (Zhang et al. 2017)	-	55.50	70.20	80.20
CDS (Tesfaye et al. 2017)	-	68.20	-	-
SFT (Zhou et al. 2017)	50.70	70.60	90.00	97.60
DCF (Li et al. 2017a)	56.05	71.77	86.57	93.08
SeeForest (Zhou et al. 2017)	50.70	70.60	90.00	97.60
DRSA (Li et al. 2018)	65.80	82.30	-	-
DuATM (Si et al. 2018)	67.73	81.16	92.47	-
LSTM (Yan et al. 2016)	61.58	76.11	85.30	92.68
A&O (Simonyan et al. 2014)	63.39	77.11	88.41	94.60
Two-stream M3D	**74.06**	**84.39**	**93.84**	**97.74**

Table 4: Comparisons on *PRID* and *iLIDS-VID*.

Dataset	*PRID*		*iLIDS-VID*	
Method	r1	r5	r1	r5
BoW+XQDA (Zheng et al. 2016)	31.80	58.50	14.00	32.20
DVDL (Karanam et al. 2015)	40.60	69.70	25.90	48.20
RFA-Net (Yan et al. 2016)	58.20	85.80	49.30	76.80
STFV3D (Koestinger et al. 2012)	64.10	87.30	44.30	71.70
DRCN (Wu et al. 2016)	69.00	88.40	46.10	76.80
RCN (McLaughlin et al. 2016)	70.00	90.00	58.00	84.00
IDE+XQDA (Zheng et al. 2016)	77.30	93.50	53.00	81.40
DFCP (Li et al. 2017b)	51.60	83.10	34.30	63.30
SeeForest (Zhou et al. 2017)	79.40	94.40	55.20	86.50
AMOC (Liu et al. 2017a)	83.70	98.30	68.70	94.30
QAN (Liu et al. 2017b)	90.30	98.20	68.00	86.80
DRSA (Li et al. 2018)	93.20	-	**80.20**	-
Two-stream M3D	**94.40**	**100.00**	74.00	**94.33**

It is clear that, any one of the 3 attention branches consistently improves the performance of M3D. Combining the complete RAL brings the most substantial performance gains. For example, RAL boosts the rank-1 accuracy of M3D from 87.64% to 91.03% on *PRID*. This demonstates the validity of our RAL in identifying discriminative spatio-temporal feature. It also shows the advantages of introducing attention mechanism in video feature learning.

3) Spatial-Temporal Feature Fusion This part tests the performance of our two-stream convolution network which involves the M3D convolution layer, RAL, and the spatial-temporal feature fusion. The comparisons on three datasets are summarized in Table 2. "Two-stream M3D" refers to the complete two-stream architecture in Fig. 3.

It can be observed from Table 2 that, M3D CNN outperforms the 2D baseline by large margins by considering extra temporal information. This shows the benefits of considering temporal cues in video based person ReID. The attention layer RAL further boosts the performance of M3D CNN. Combining the 2D CNN and M3D CNN features achieves the best performance in Table 2. This shows our two-stream architecture is effective to exploit the complementary information cross spatial and temporal domains. In the following part, we compare our two-stream M3D network with recent works on three datasets.

Comparison with Recent Work

Table 3 reports the comparison of our approach with recent works on *MARS*. It can be observed from Table 3 that, our method constantly outperforms all of the compared methods. Our method achieves the rank1 accuracy of 84.39% and mAP of 74.06%, outperforming two latest works DuATM (Si et al. 2018) and DRSA (Li et al. 2018) by 6.33% and 8.26% in mAP, respectively. Note that, DRSA (Li et al. 2018) extracts local part features to gain stronger discriminative power. DuATM (Si et al. 2018) introduces a complex frame feature matching strategy with quadratic complexity. Compared with those two works, our method is more concise and efficient, *e.g.*, we extract global feature and match features with simple Euclidean distance.

We further build two widely used temporal feature extraction methods based on ResNet50, *e.g.*, LSTM (Yan et al. 2016) and Appearance&Optical flow (Simonyan and Zisserman 2014). The comparison in Table 3 clearly shows that, our method outperforms those temporal feature extraction works. For example, our method outperforms the LSTM based method by 12.48% in mAP. This significant performance boost demonstrates the advantage of our two-stream M3D network in spatial-temporal feature learning.

The comparisons on *PRID* and *iLIDS-VID* datasets are shown in Table 4. As shown in the table, our proposed method presents competitive performance on rank1 accuracy. DRSA (Li et al. 2018) also gets competitive performance on both datasets, and outperforms our method on *iLIDS-VID* dataset. The reason may be because *iLIDS-VID* has a small training set. DRSA alleviates the insufficiency of training data using multi-task learning strategy on part cues. DRSA also impose Online Instance Matching loss (OIM) loss for training, which is shown more effective than our softmax. Extracting global feature trained with basic softmax loss, our method still outperforms DRSA on the other two datasets. Our competitive performance demonstrates the advantage of learning spatial-temporal cues in person ReID.

Conclusion

This paper proposes a two-stream convolution network to explicitly leverages spatial and temporal cues for video based person ReID. A novel Multi-scale 3D (M3D) network is constructed to learn the multi-scale temporal cues in video sequences. Implemented by inserting serval M3D convolution layers into 2D CNN networks, M3D network can learn robust temporal representations with a fraction of increased parameters. A Residual Attention Layer (RAL) is further designed to refine the learned temporal features by M3D in residual manner. The learned temporal representations are combined with spatial representation learned through 2D CNN for video ReID. Experimental results on three widely used video ReID datasets demonstrate the superiority of the proposed model over current state-of-the-art methods.

Acknowledgments This work is supported in part by Peng Cheng Laboratory, in part by Beijing Natural Science Foundation under Grant No. JQ18012, in part by Natural Science Foundation of China under Grant No. 61620106009, 61572050, 91538111.

References

Carreira, J., and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.

Feichtenhofer, C.; Pinz, A.; and Wildes, R. P. 2017. Spatiotemporal multiplier networks for video action recognition. In *CVPR*.

Feichtenhofer, C.; Pinz, A.; and Zisserman, A. 2016. Convolutional two-stream network fusion for video action recognition. In *CVPR*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

Hirzer, M.; Beleznai, C.; Roth, P. M.; and Bischof, H. 2011. Person re-identification by descriptive and discriminative classification. In *Image Analysis*.

Hu, J.; Shen, L.; and Sun, G. 2017. Squeeze-and-excitation networks. *arXiv preprint arXiv:1709.01507*.

Ji, S.; Xu, W.; Yang, M.; and Yu, K. 2013. 3d convolutional neural networks for human action recognition. *IEEE Trans. PAMI*.

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

Klaser, A.; Marszałek, M.; and Schmid, C. A spatio-temporal descriptor based on 3d-gradients. In *BMVC*.

Koestinger, M.; Hirzer, M.; Wohlhart, P.; Roth, P. M.; and Bischof, H. 2012. Large scale metric learning from equivalence constraints. In *CVPR*.

Li, D.; Chen, X.; Zhang, Z.; and Huang, K. 2017a. Learning deep context-aware features over body and latent parts for person re-identification. In *CVPR*.

Li, Y.; Zhuo, L.; Li, J.; Zhang, J.; Liang, X.; and Tian, Q. 2017b. Video-based person re-identification by deep feature guided pooling. In *CVPR Workshops*.

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

Li, M.; Zhu, X.; and Gong, S. 2018a. Unsupervised person re-identification by deep learning tracklet association. In *ECCV*.

Li, W.; Zhu, X.; and Gong, S. 2018b. Harmonious attention network for person re-identification. In *CVPR*.

Liu, H.; Jie, Z.; Jayashree, K.; Qi, M.; Jiang, J.; Yan, S.; and Feng, J. 2017a. Video-based person re-identification with accumulative motion context. *IEEE Trans. CSVT*.

Liu, Y.; Yan, J.; Ouyang, W.; and Ouyang, W. 2017b. Quality aware network for set to set recognition. In *CVPR*.

Liu, J.; Zha, Z.-J.; Chen, X.; Wang, Z.; and Zhang, Y. 2018a. Dense 3d-convolutional neural network for person re-identification in videos. *ACM Trans. Multimed. Comput. Commun. Appl.*

Liu, J.; Zha, Z.-J.; Xie, H.; Xiong, Z.; and Zhang, Y. 2018b. Ca 3 net: Contextual-attentional attribute-appearance network for person re-identification. In *ACM MM*.

McLaughlin, N.; Martinez del Rincon, J.; Miller, P.; and Miller, P. 2016. Recurrent convolutional network for video-based person re-identification. In *CVPR*.

Pedagadi, S.; Orwell, J.; Velastin, S.; and Boghossian, B. 2013. Local fisher discriminant analysis for pedestrian re-identification. In *CVPR*.

Qiu, Z.; Yao, T.; and Mei, T. 2017. Learning spatio-temporal representation with pseudo-3d residual networks. In *ICCV*.

Scovanner, P.; Ali, S.; and Shah, M. 2007. A 3-dimensional sift descriptor and its application to action recognition. In *ACM MM*.

Si, J.; Zhang, H.; Li, C.-G.; Kuen, J.; Kong, X.; Kot, A. C.; and Wang, G. 2018. Dual attention matching network for context-aware feature sequence based person re-identification. *arXiv preprint arXiv:1803.09937*.

Simonyan, K., and Zisserman, A. 2014. Two-stream convolutional networks for action recognition in videos. In *NIPS*.

Su, C.; Yang, F.; Zhang, S.; Tian, Q.; Davis, L. S.; and Gao, W. 2015. Multi-task learning with low rank attribute embedding for person re-identification. In *ICCV*.

Su, C.; Zhang, S.; Xing, J.; Gao, W.; and Tian, Q. 2016. Deep attributes driven multi-camera person re-identification. In *ECCV*. Springer.

Su, C.; Li, J.; Zhang, S.; Xing, J.; Gao, W.; and Tian, Q. 2017. Pose-driven deep convolutional model for person re-identification. In *ICCV*.

Tesfaye, Y. T.; Zemene, E.; Prati, A.; Pelillo, M.; and Shah, M. 2017. Multi-target tracking in multiple non-overlapping cameras using constrained dominant sets. *arXiv preprint arXiv:1706.06196*.

Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*.

Tran, D.; Wang, H.; Torresani, L.; Ray, J.; LeCun, Y.; and Paluri, M. 2018. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*.

Wang, X., and Zhao, R. 2014. Person re-identification: System design and evaluation overview. In *Person Re-Identification*.

Wei, L.; Zhang, S.; Yao, H.; Gao, W.; and Tian, Q. 2017. Glad: global-local-alignment descriptor for pedestrian retrieval. In *ACM MM*. ACM.

Wu, L.; Shen, C.; Hengel, A. v. d.; and Hengel, A. v. d. 2016. Deep recurrent convolutional networks for video-based person re-identification: An end-to-end approach. *arXiv preprint arXiv:1606.01609*.

Wu, Y.; Lin, Y.; Dong, X.; Yan, Y.; Ouyang, W.; and Yang, Y. 2018. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In *CVPR*.

Xiong, F.; Gou, M.; Camps, O.; and Sznaier, M. 2014. Person re-identification using kernel-based metric learning methods. In *ECCV*.

Xu, J.; Zhao, R.; Zhu, F.; Wang, H.; and Ouyang, W. 2018. Attention-aware compositional network for person re-identification. *CVPR*.

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

Ye, M.; Lan, X.; and Yuen, P. C. 2018. Robust anchor embedding for unsupervised video person re-identification in the wild. In *ECCV*.

Yu, F., and Koltun, V. 2015. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*.

Zhang, W.; Hu, S.; Liu, K.; and Liu, K. 2017. Learning compact appearance representation for video-based person re-identification. *arXiv preprint arXiv:1702.06294*.

Zheng, L.; Bie, Z.; Sun, Y.; Wang, J.; Su, C.; Wang, S.; and Tian, Q. 2016. Mars: A video benchmark for large-scale person re-identification. In *ECCV*.

Zhou, Z.; Huang, Y.; Wang, W.; Wang, L.; and Tan, T. 2017. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In *CVPR*.