北京大学数字媒体研究所
INSTITUTE OF DIGITAL MEDIA, PEKING UNIVERSITY

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, China

2019/01/28

Outline

- Background
- Our Approach
- Experiment
- Take home message

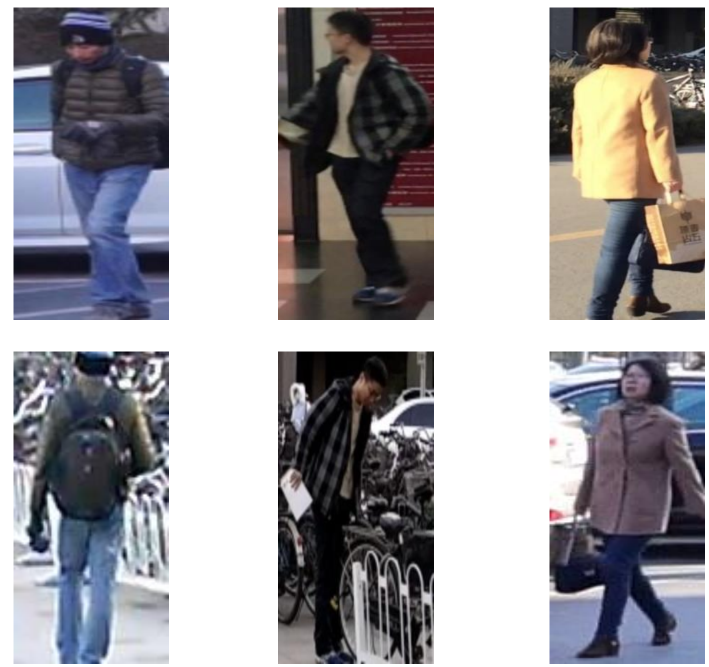
Outline

- Background
- Our Approach
- Experiment
- Take home message

Problem Statement

- In non-overlapping camera networks, matching the same individuals across multiple cameras.
- Person ReID has many challenging issues like:
  - Viewpoint change
  - Lighting change
  - Pose change

Large Intra-class Variation
Small Inter-class Variation
Temporal cues maybe more useful !

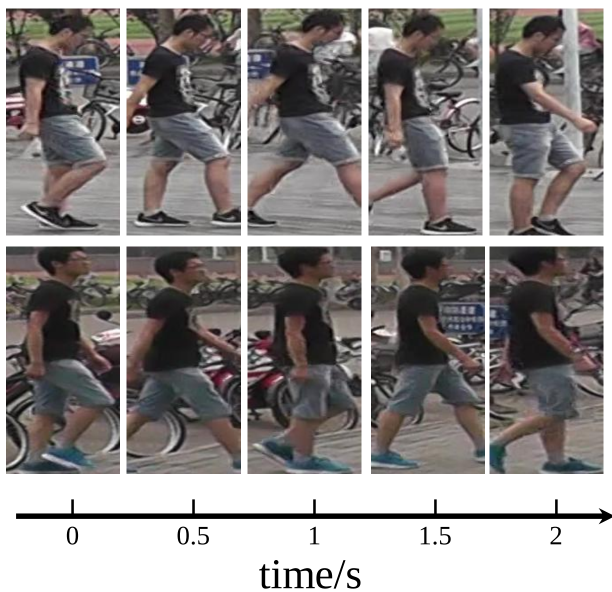


Pose **viewpoint** **Lighting**

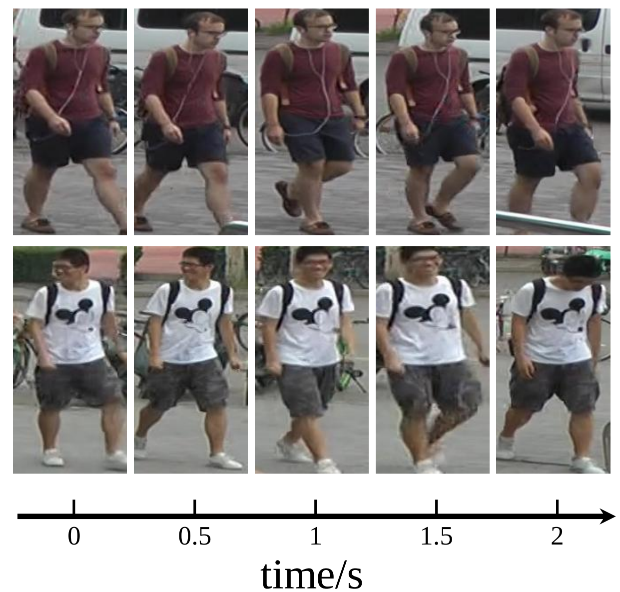
Motivation

- Temporal cues are equally important with spatial.

Similar appearance
Different gait



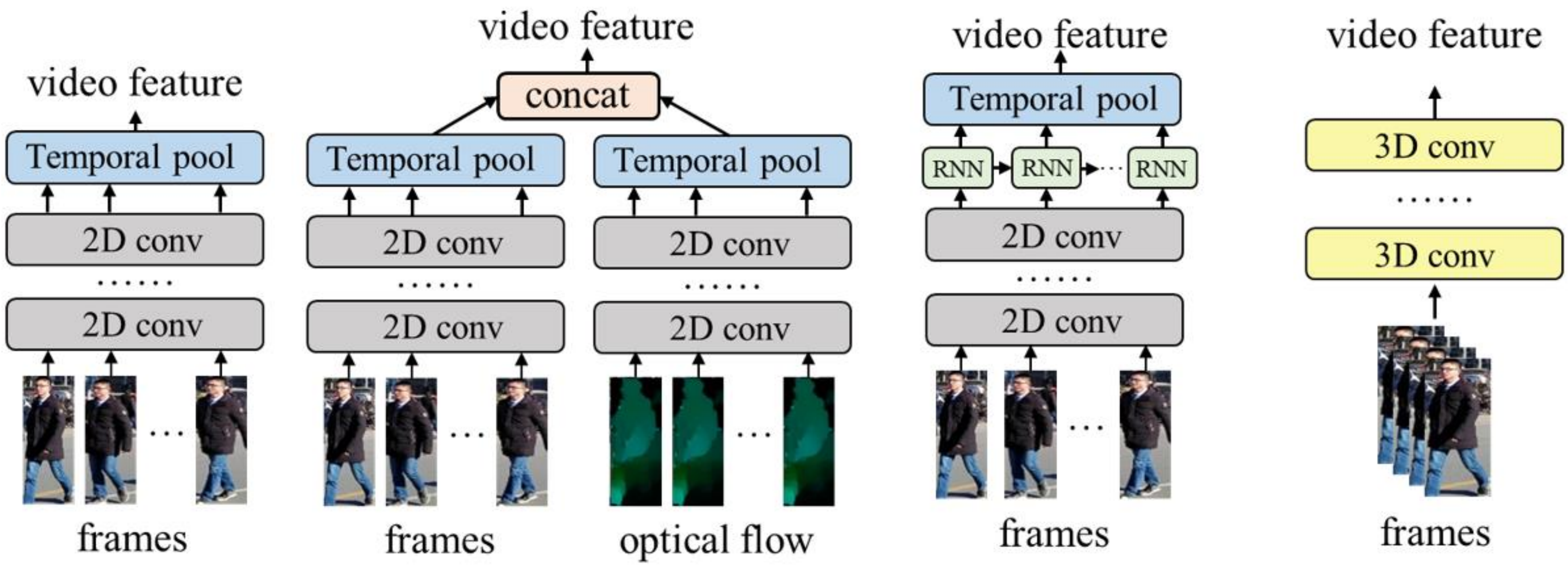
Similar gait
Different appearance



Leverage temporal information is important

Existing temporal methods

- Existing temporal feature learning methods:



(a) 2D CNN [zheng, ECCV16]

(b) Two-Stream CNN [Chung, ICCV17]

(c) RNN [McLaughlin, CVPR16]

(d) 3D CNN [Carreira, CVPR17]

Motivation

- Occlusion is unavoidable in real scene, which lead to low quality frames.

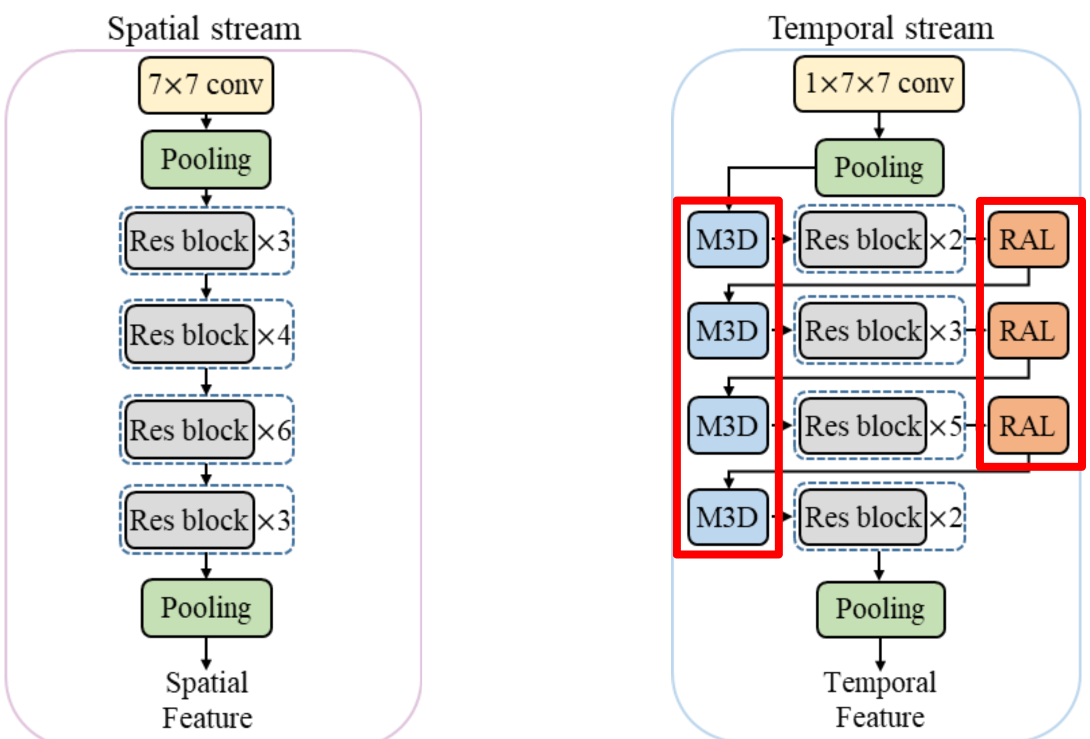
How to relieve the influence of low quality frames?

Outline

- Background
- Our Approach
- Experiment
- Take home message

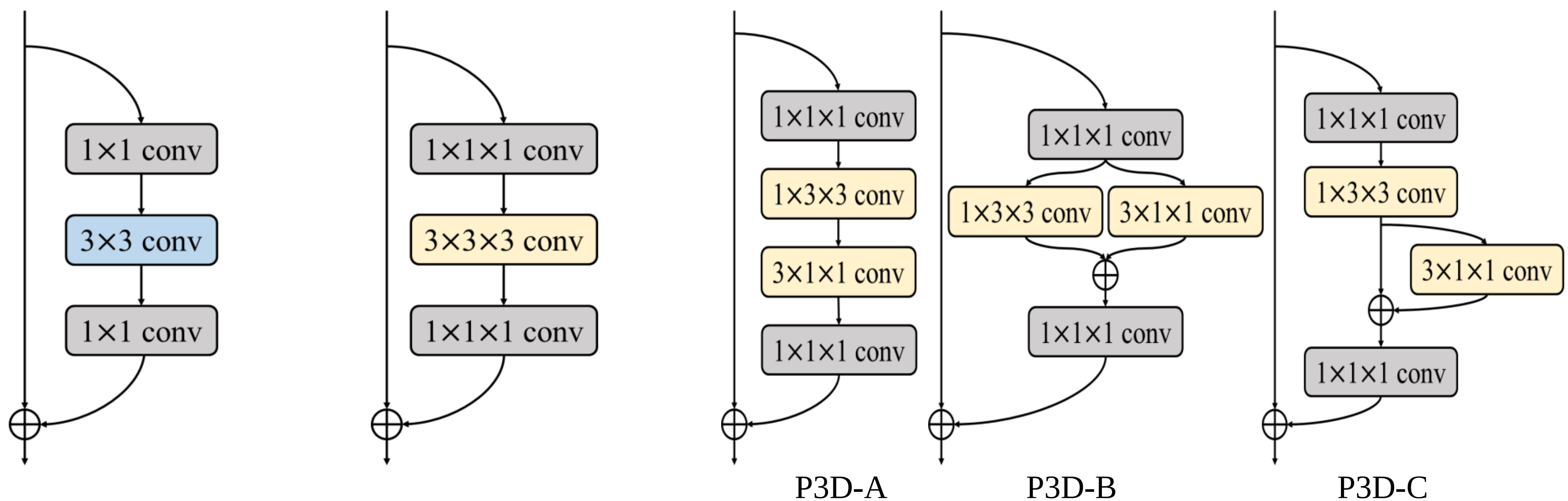
Two-stream Network

- 2D CNN for spatial feature learning
- M3D and RAL layers are inserted into 2D CNN for temporal feature learning



Shortcoming of existing 3D CNN

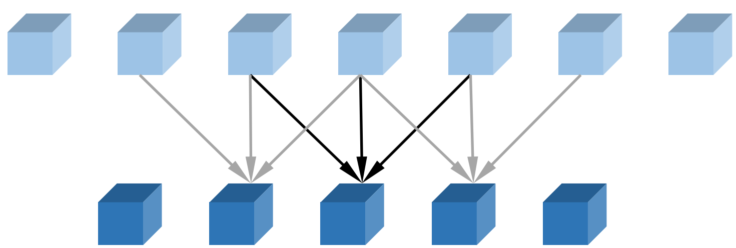
- Small receptive field
- Large number of parameters
- Can’t fully utilize ImageNet pre-trained model



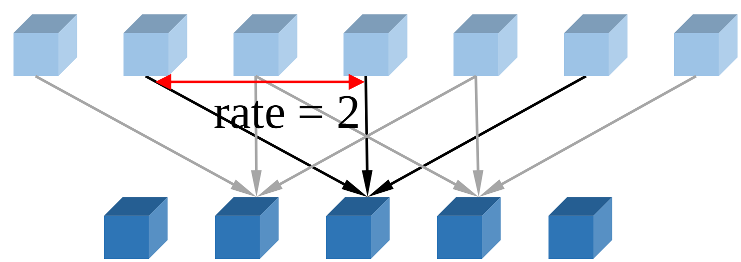
2D CNN[Zheng, ECCV16] I3D [Carreira, CVPR17] P3D [Qiu, ICCV17]

Basic idea

- Dilation convolution has same number of parameters, but larger receptive field
- Impose parallel dilation convolutions can jointly learn multi-scale cues.



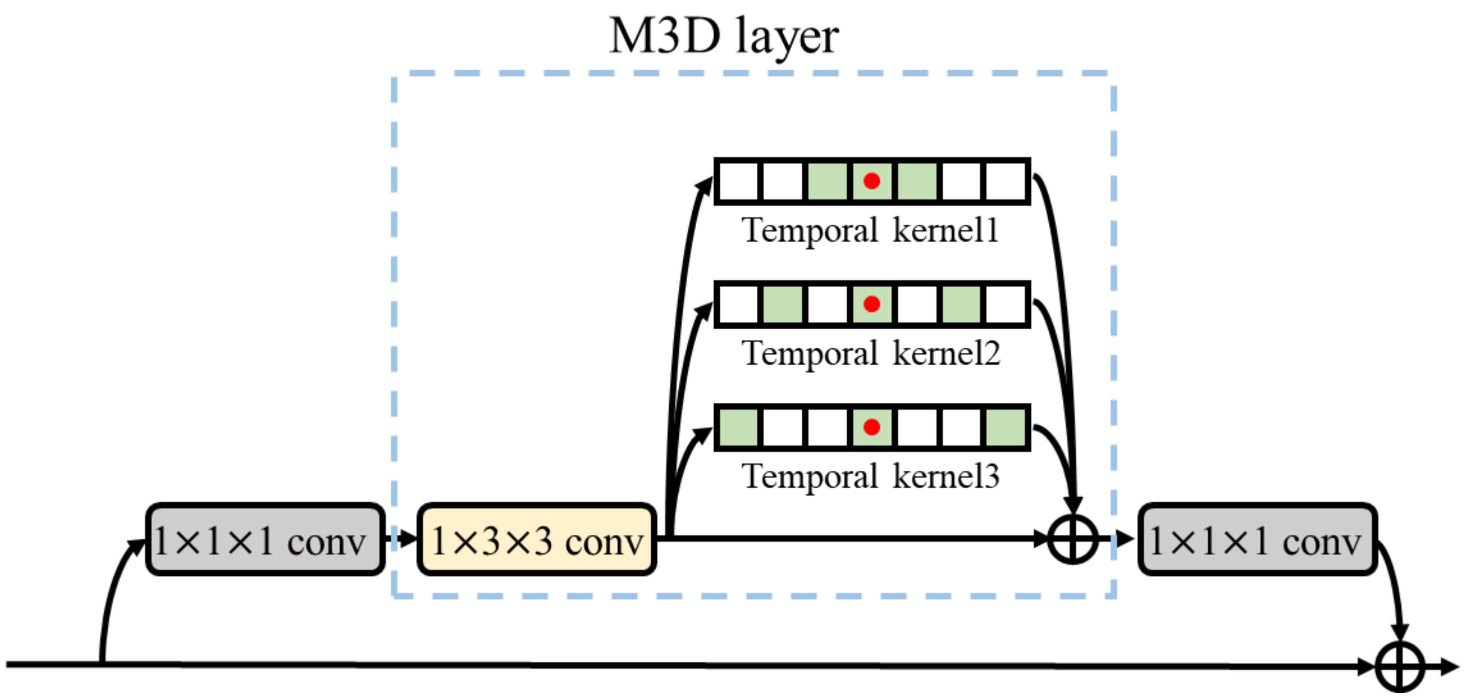
(1) Standard Convolution



(2) Dilated Convolution [Yu, ICLR16]

The Multi-scale 3D Convolution

- Multi-scale receptive field
- Less parameters
- Take advantage of 2D pre-trained model



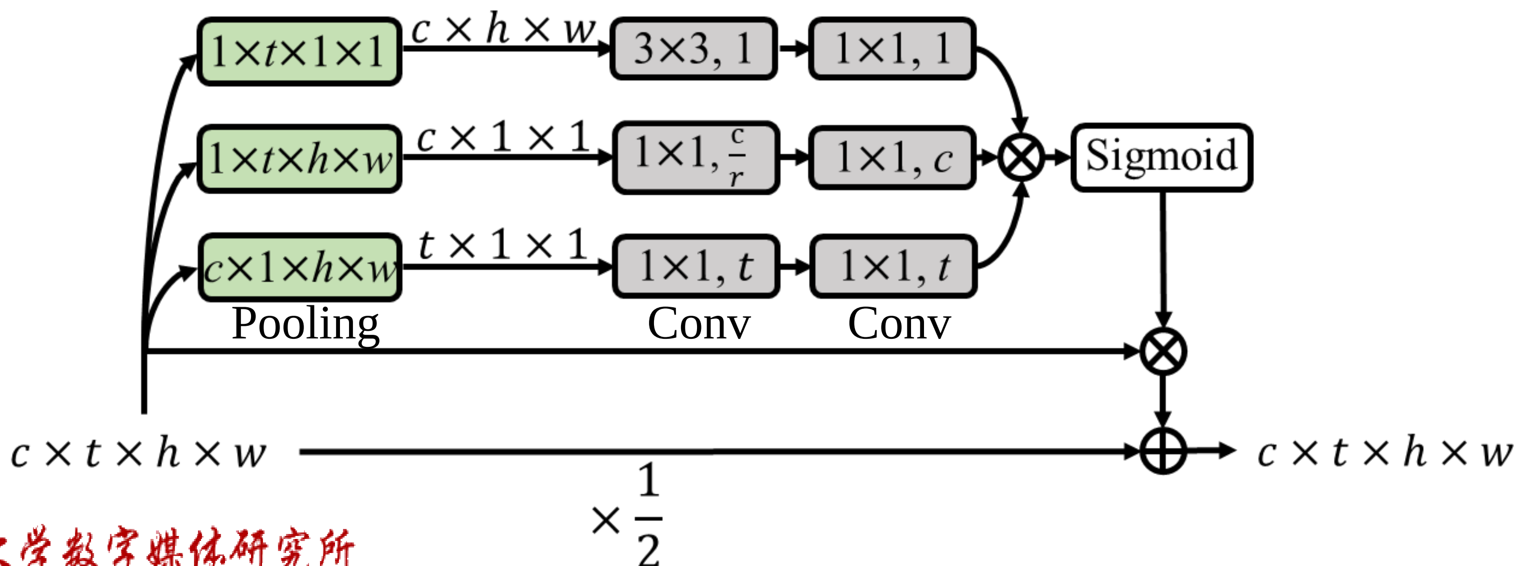
The Residual Attention Layer

- Decompose attention learning into three branches:

$$M = Sigmoid(S_m \times C_m \times T_m)$$

- The attention is residual connection to keep original initialization manner:

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



Summary

- Propose a novel M3D layer to learn multi-scale temporal cues
- Propose RAL to relieve the influence of low quality frame
- Introduce two-stream architecture for spatial temporal feature learning

Outline

- Background
- Our Approach
- Experiment
- Take home message

Evaluation protocols

- We select three video ReID datasets as our evaluation protocols, including:
  - **PRID-2011** : 400 sequences of 200 pedestrians under 2 cameras
  - **iLIDS-VID** : 600 sequences of 300 pedestrians under 2 cameras
  - **MARS** : 1261 pedestrians and 20,715 sequences under 6 cameras

PRID-2011

iLIDS-VID

MARS

Comparison with 3D CNNs

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	

Better performance !

Less parameter !

Higher speed !

Effectiveness of each component

- Consider all components, the two-stream get best performance.

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**

Comparison on MARS

MARS	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**

Comparison with recent work

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**

Examples of ReID result on MARS

Query:

True match

Ours:

Baseline :

Examples of ReID result on MARS

True match

Query:

Ours:

Baseline :

Outline

- Background
- Our Approach
- Experiment
- Take home message

Take home message

- New 3D CNN is proposed with
  - Less parameters and fast speed
  - Capture multi-scale temporal cues
  - Easy to train
- The proposed two-stream M3D architecture shows promising performance on widely used ReID benchmarks
- Other video tasks like action recognition will be further tested.

Q&A

Thank You!

The source code have been released