Gated Deep Neural Networks for Adaptive Information Flow

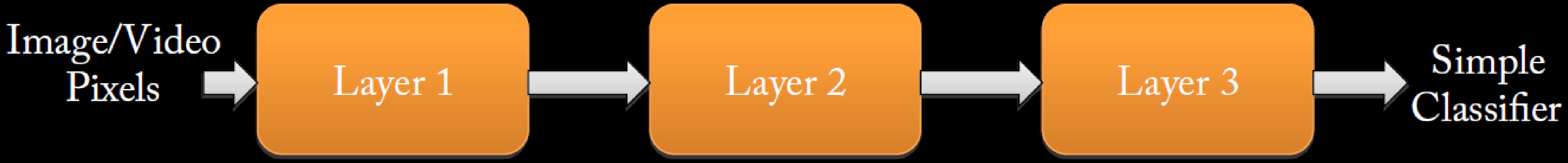
Wang Gang

Alibaba Group

Ali A.I. Labs
阿里人工智能实验室

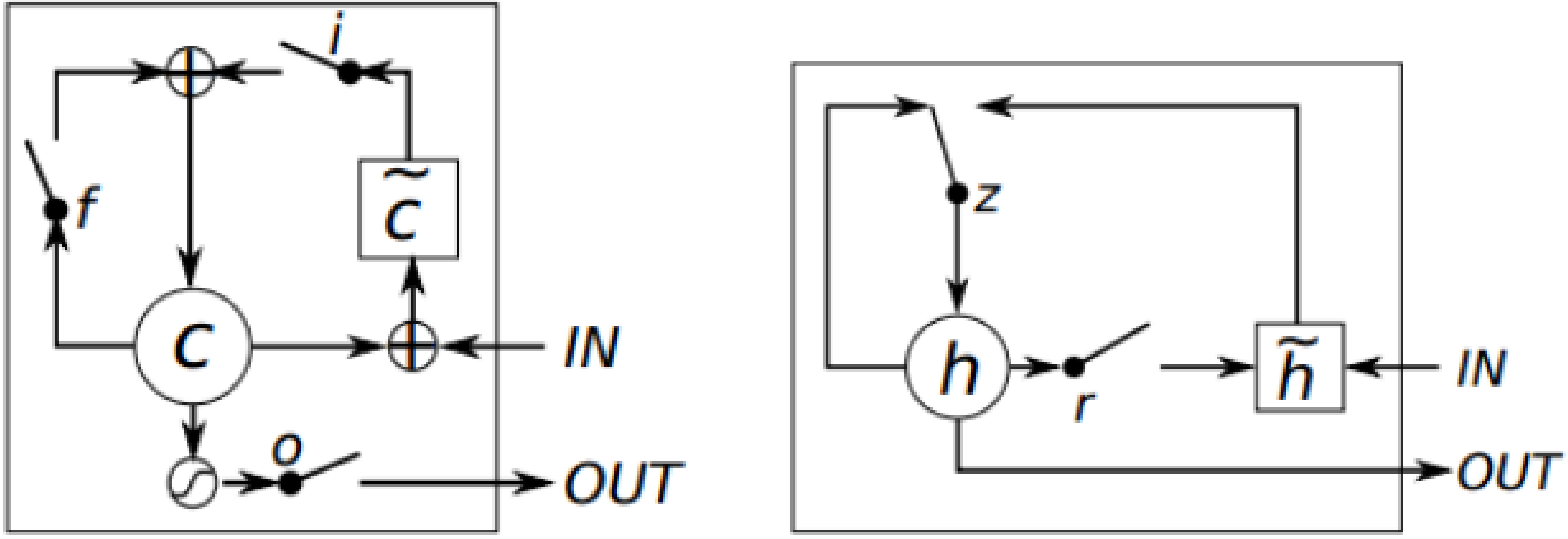
- AI Labs was established in May 2016, under the theme of Alibaba’s New Technology.
- We are devoted to developing advanced technologies on computer vision, natural language processing, and interaction, and we encourage publishing.
- We are developing our own Artificial Intelligent products for consumers and business partners.

Problem of Inference Process



- We extract features/information from each layer and propagate to the next layer in the deep learning networks.
- Currently, most existing deep neural networks have rigid information propagation structure in the forward process.

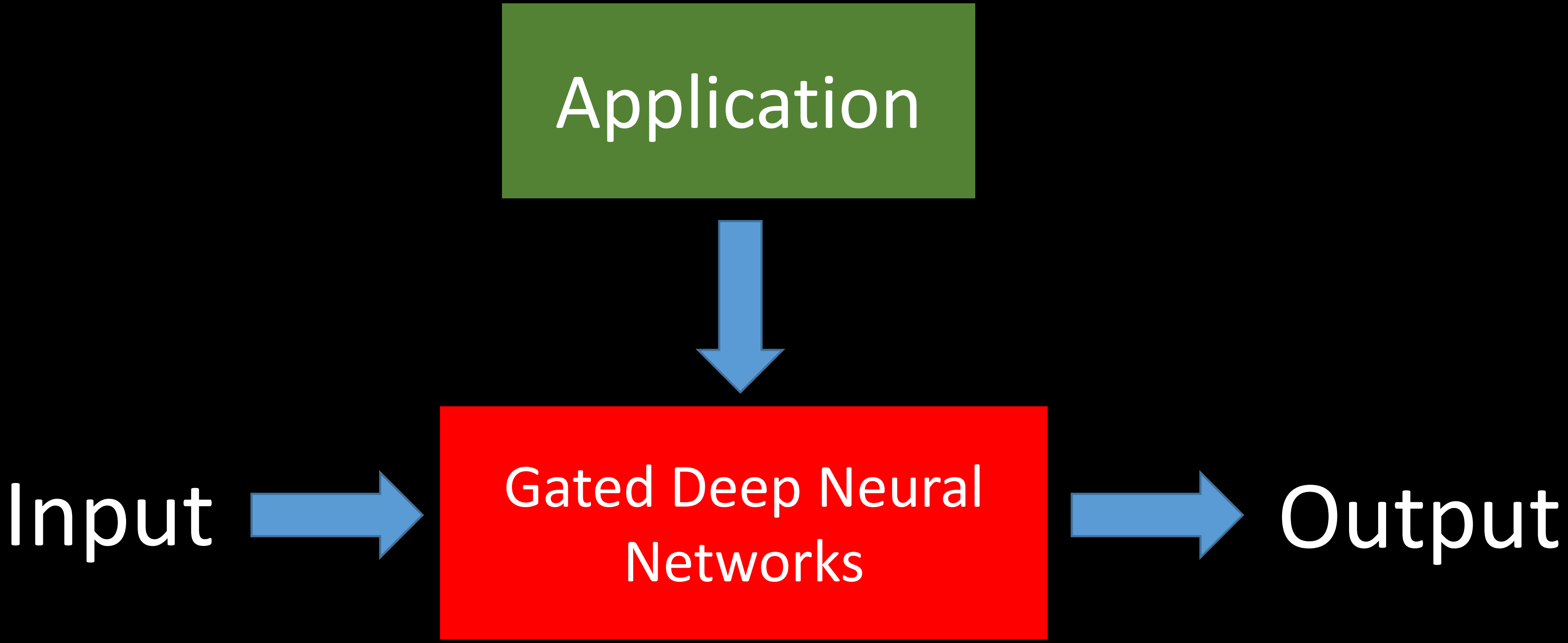
Successful Attempts using Gating Function



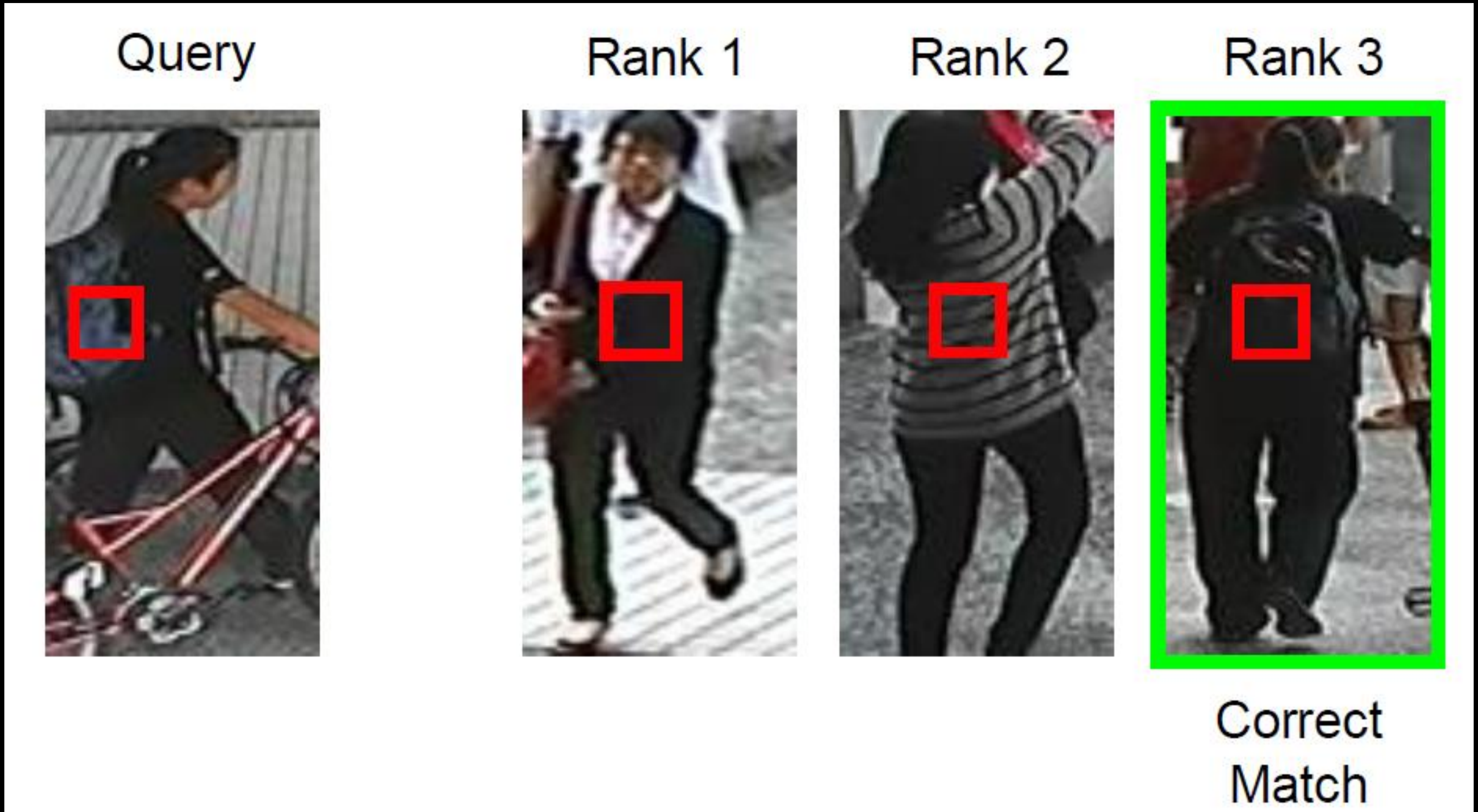
(a) Long Short-Term Memory

(b) Gated Recurrent Unit

Gated Deep Neural Networks for Specific Applications

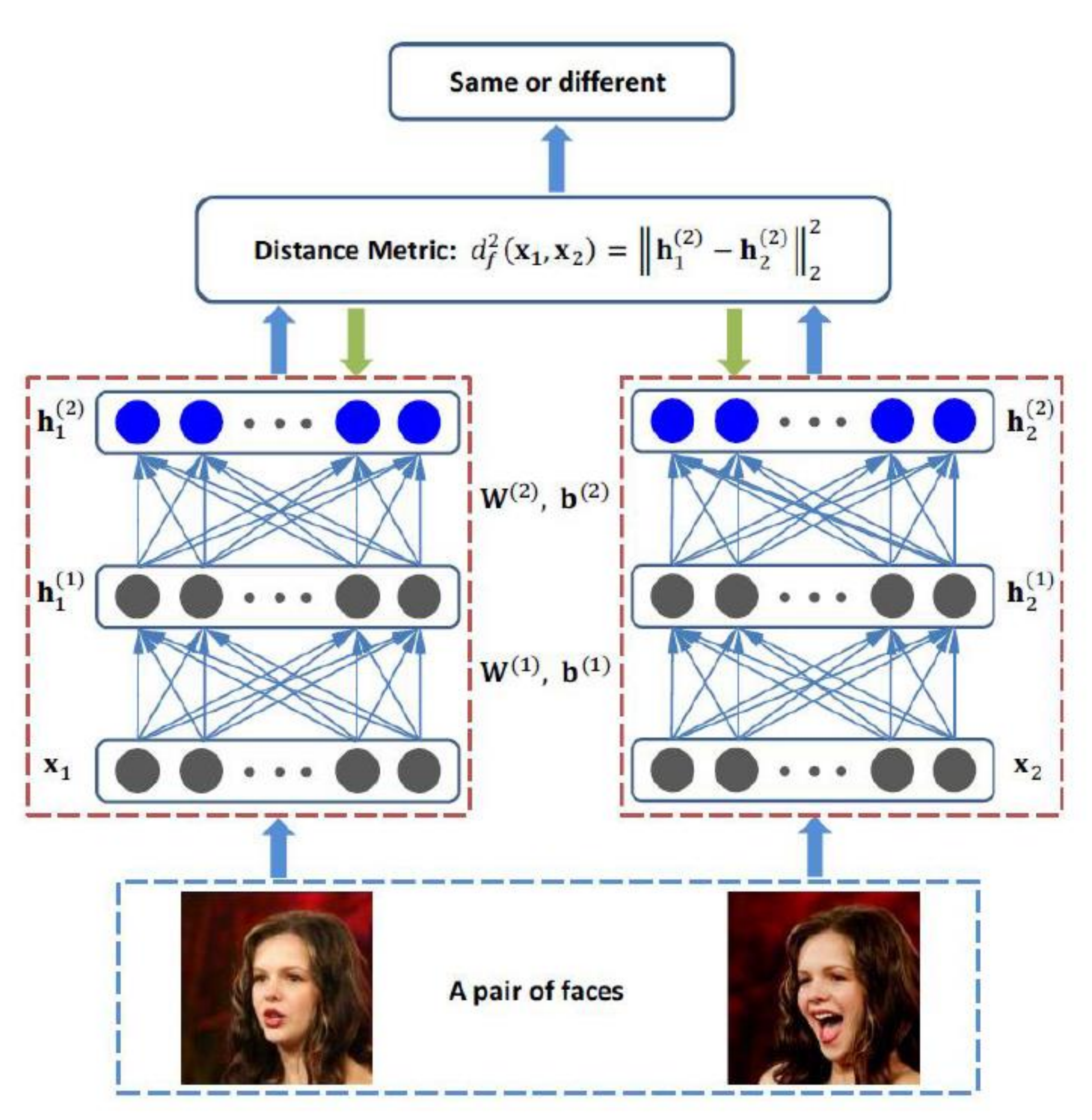


Matching Gates for Human Re-identification



Rahul Rama Varior, Mrinal Haloi, Gang Wang, "Gated Siamese Convolutional Neural Network Architecture for Human Re-Identification," ECCV, 2016

Conventional Siamese Neural Networks


Same or different
Distance Metric: $d_f^2(\mathbf{x}_1, \mathbf{x}_2) = \left\| \mathbf{h}_1^{(2)} - \mathbf{h}_2^{(2)} \right\|_2^2$
$\mathbf{h}_1^{(2)}$
$\mathbf{h}_2^{(2)}$
$\mathbf{W}^{(2)}, \mathbf{b}^{(2)}$
$\mathbf{h}_1^{(1)}$
$\mathbf{h}_2^{(1)}$
$\mathbf{W}^{(1)}, \mathbf{b}^{(1)}$
$\mathbf{x}_1$
$\mathbf{x}_2$
A pair of faces

- Siamese networks share the same parameters to extract features from pairs of images for similarity comparison
- Such networks achieved state-of-the-art performance on many similarity comparison tasks such as human re-identification and face verification
- When extracting features from one pair of images, the networks are not aware of the other images, and may ignore important similarity patterns for this pair at the middle layers

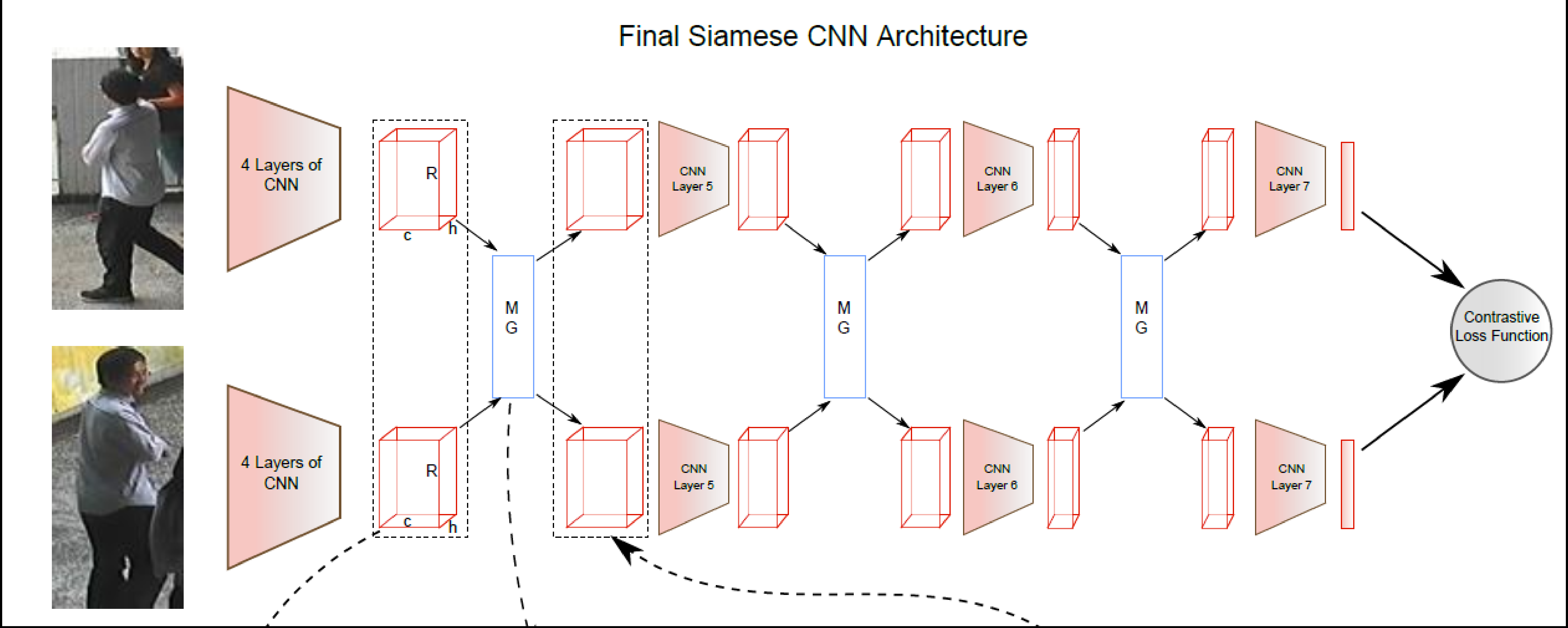
Our Proposal: Matching Gates

- A gating function to compare the extracted local patterns for an image pair starting from the mid-level and promote (i.e. to amplify) the local similar patterns so that the network propagates more relevant features to the higher layers of the network.

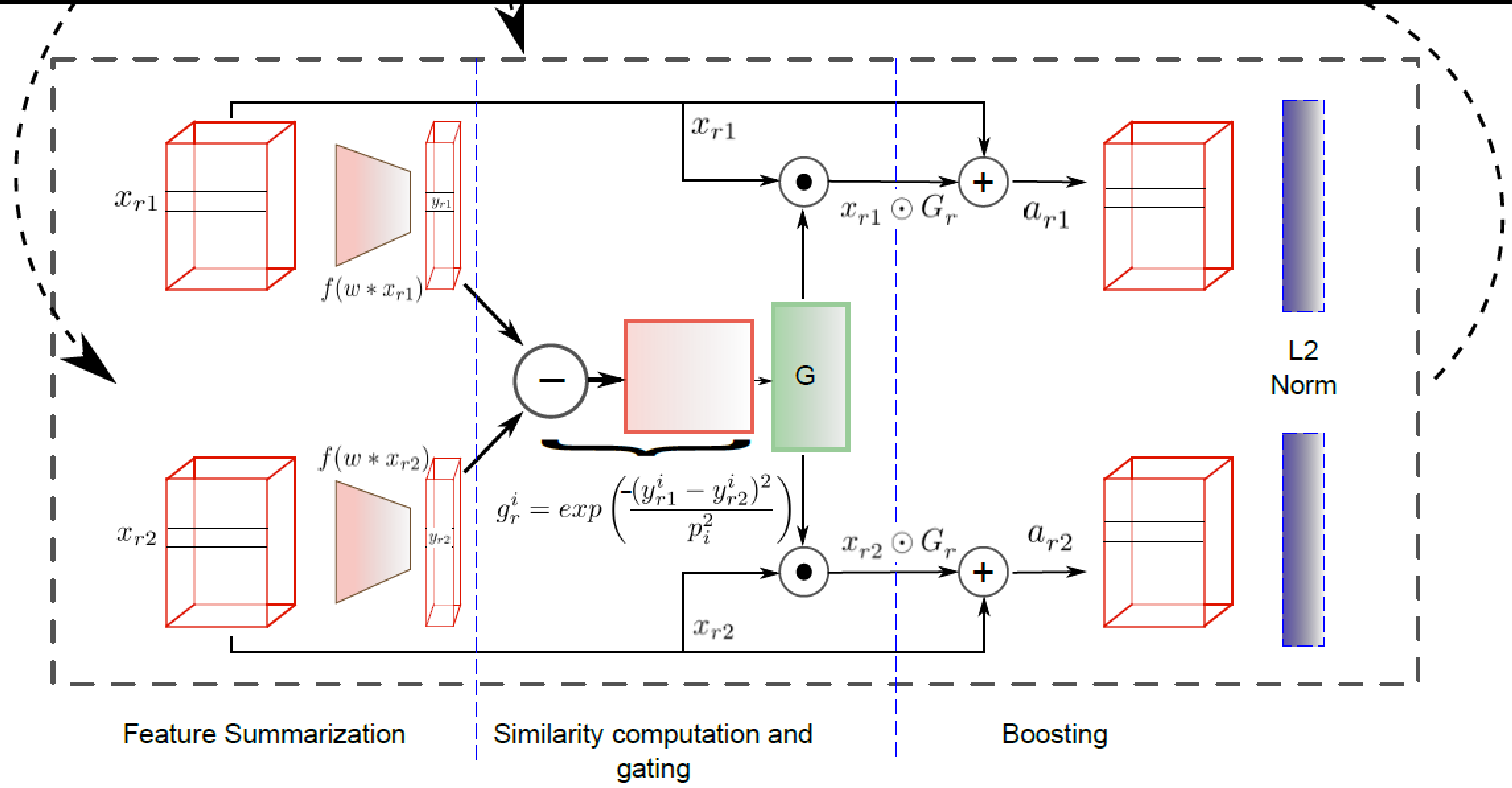
- During training phase, the mechanisms inside the gating function also boosts the back propagated gradients corresponding to the amplified local similarities, to learn patterns which are more locally similar.

- A way of controlling the information flow to be adaptive to the application using gating functions.

Our Framework



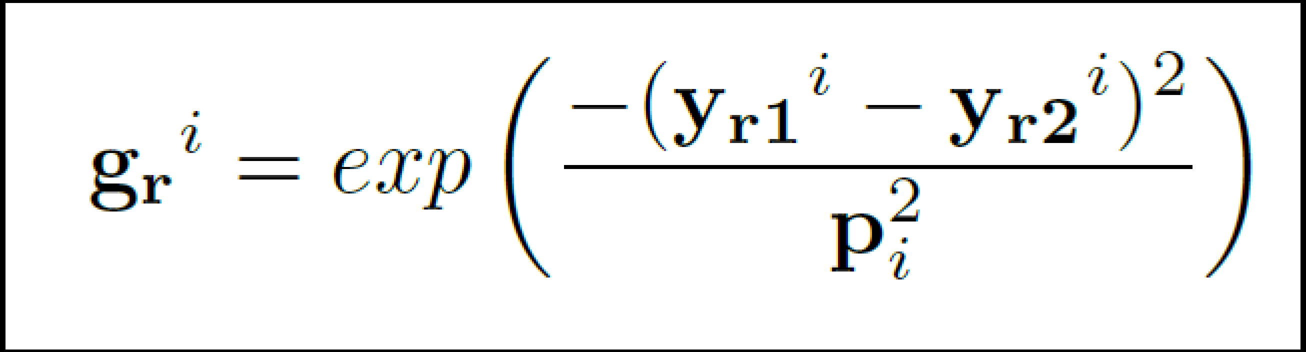
The Matching Gate



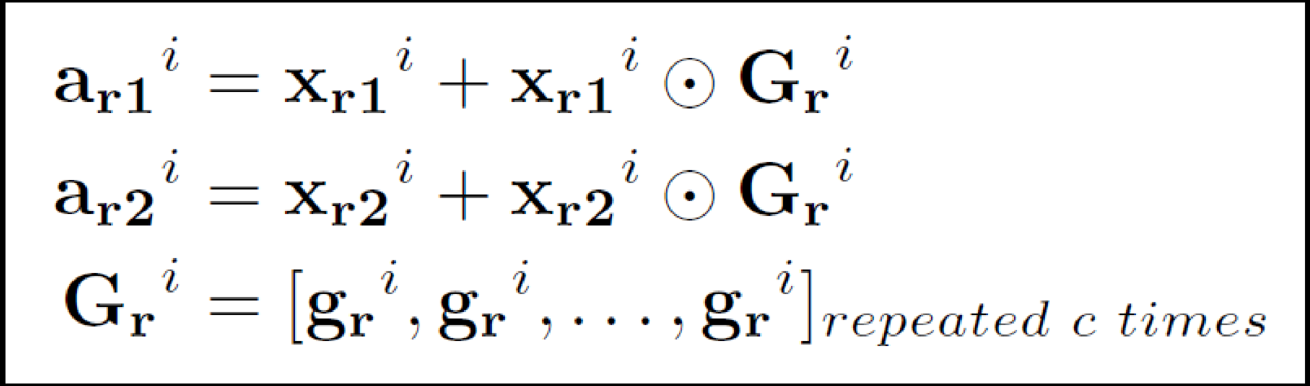
Matching Gate Architecture

M.G :- Matching Gate

Formulation



$$\mathbf{g_r}^i = exp\left(\frac{-(\mathbf{y_{r1}}^i - \mathbf{y_{r2}}^i)^2}{\mathbf{p}_i^2}\right)$$



$$\mathbf{a_{r1}}^i = \mathbf{x_{r1}}^i + \mathbf{x_{r1}}^i \odot \mathbf{G_r}^i$$

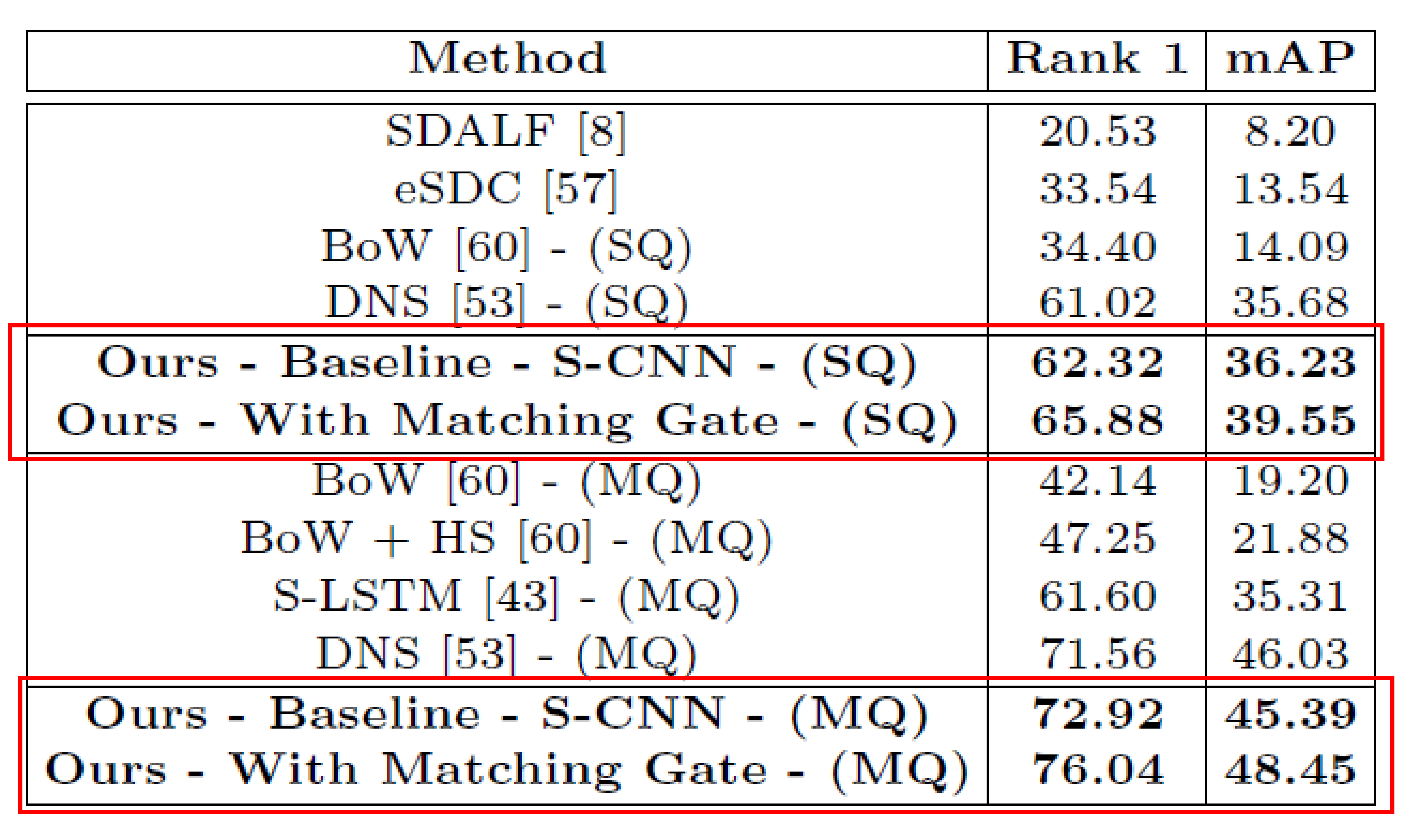
$$\mathbf{a_{r2}}^i = \mathbf{x_{r2}}^i + \mathbf{x_{r2}}^i \odot \mathbf{G_r}^i$$

$$\mathbf{G_r}^i = [\mathbf{g_r}^i, \mathbf{g_r}^i, \ldots, \mathbf{g_r}^i]_{repeated\ c\ times}$$

Experiment

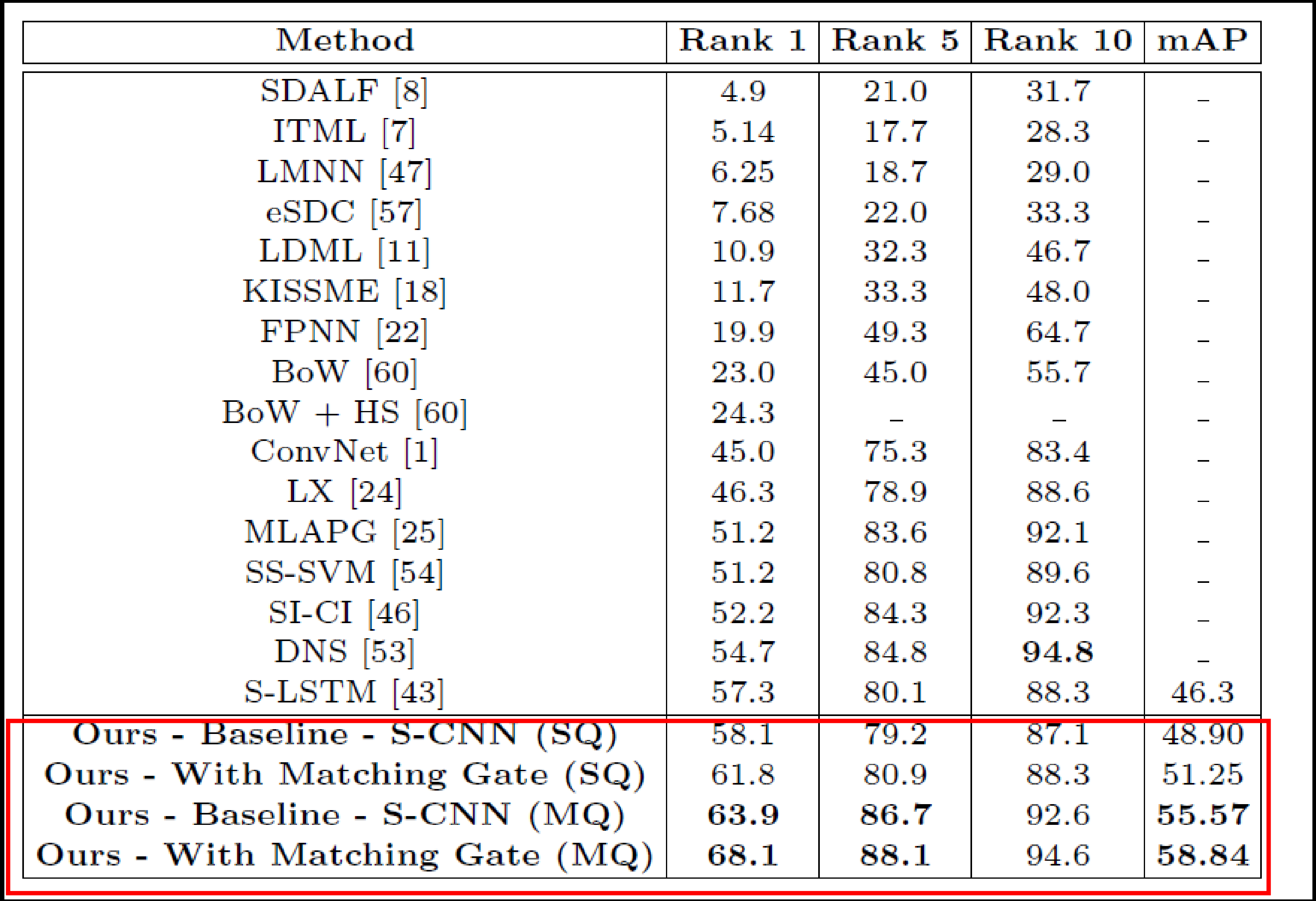
- Siamese networks take image pairs as inputs with 1/0 as the labels.
- Market-1501: The Market-1501 dataset contains 32668 annotated bounding boxes of 1501 subjects captured from 6 cameras and is currently the largest dataset for human re-identification.
- CUHK03: CUHK03 dataset contains 13164 images of 1360 subjects collected on the CUHK campus.
- Human re-identification can be treated as a retrieval problem and the mean average precision (mAP) is also reported along with the Rank- 1 accuracy (R1 Acc).

Results on the Market Dataset



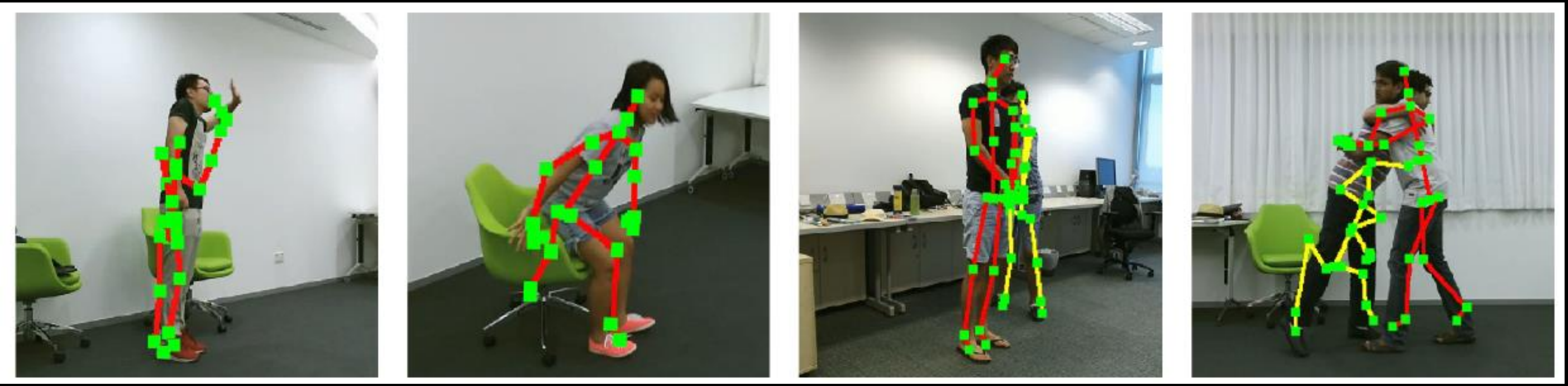
Method	Rank 1	mAP
SDALF [8]	20.53	8.20
eSDC [57]	33.54	13.54
BoW [60] - (SQ)	34.40	14.09
DNS [53] - (SQ)	61.02	35.68
Ours - Baseline - S-CNN - (SQ)	**62.32**	**36.23**
Ours - With Matching Gate - (SQ)	**65.88**	**39.55**
BoW [60] - (MQ)	42.14	19.20
BoW + HS [60] - (MQ)	47.25	21.88
S-LSTM [43] - (MQ)	61.60	35.31
DNS [53] - (MQ)	71.56	46.03
Ours - Baseline - S-CNN - (MQ)	**72.92**	**45.39**
Ours - With Matching Gate - (MQ)	**76.04**	**48.45**

Results on the CUHK dataset



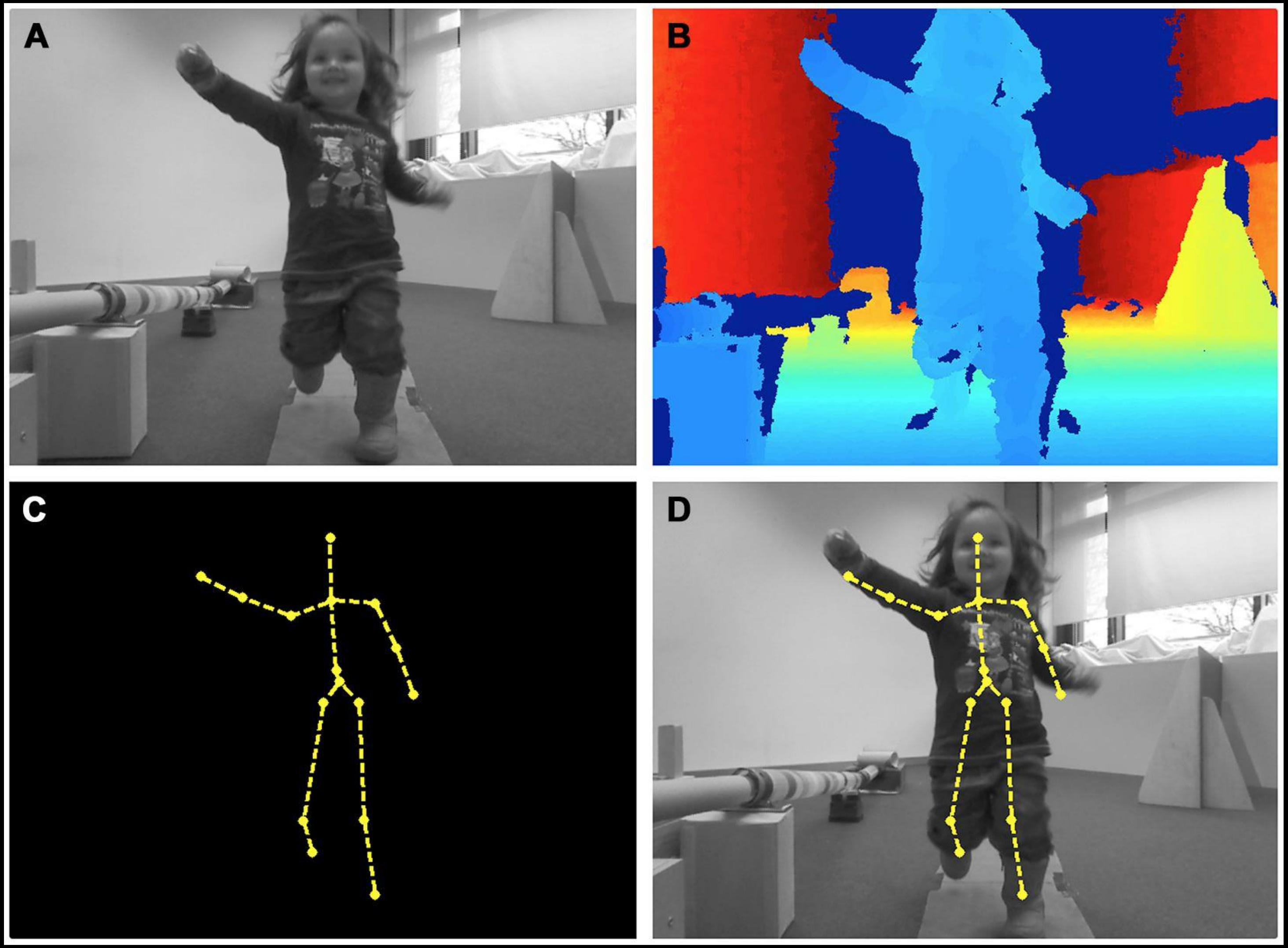
Method	**Rank 1**	**Rank 5**	**Rank 10**	**mAP**
SDALF [8]	4.9	21.0	31.7	_
ITML [7]	5.14	17.7	28.3	_
LMNN [47]	6.25	18.7	29.0	_
eSDC [57]	7.68	22.0	33.3	_
LDML [11]	10.9	32.3	46.7	_
KISSME [18]	11.7	33.3	48.0	_
FPNN [22]	19.9	49.3	64.7	_
BoW [60]	23.0	45.0	55.7	_
BoW + HS [60]	24.3	_	_	_
ConvNet [1]	45.0	75.3	83.4	_
LX [24]	46.3	78.9	88.6	_
MLAPG [25]	51.2	83.6	92.1	_
SS-SVM [54]	51.2	80.8	89.6	_
SI-CI [46]	52.2	84.3	92.3	_
DNS [53]	54.7	84.8	**94.8**	_
S-LSTM [43]	57.3	80.1	88.3	46.3
Ours - Baseline - S-CNN (SQ)	58.1	79.2	87.1	48.90
Ours - With Matching Gate (SQ)	61.8	80.9	88.3	51.25
Ours - Baseline - S-CNN (MQ)	**63.9**	**86.7**	92.6	**55.57**
Ours - With Matching Gate (MQ)	**68.1**	**88.1**	94.6	**58.84**

Trust Gates for Action Recognition



Jun Liu, Amir Shahroudy, Dong Xu, Gang Wang, "Spatio-Temporal LSTM with Trust Gates for 3D Human Action Recognition," ECCV, 2016.

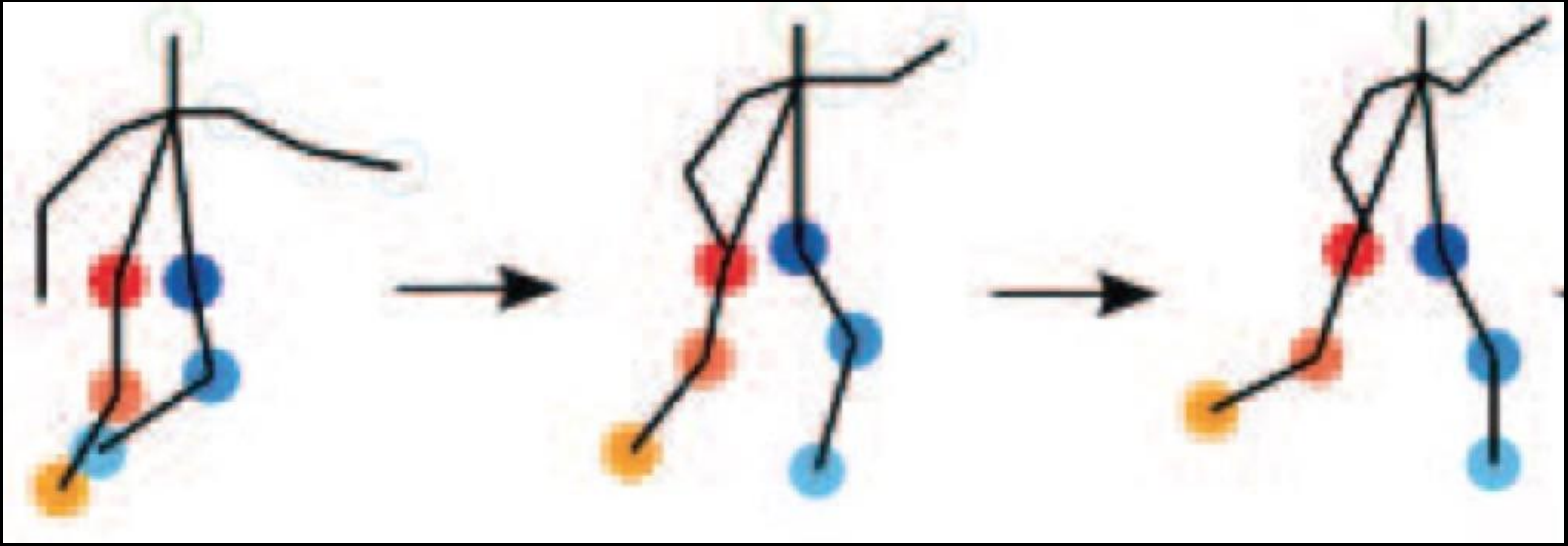
Skeleton Based Action Recognition



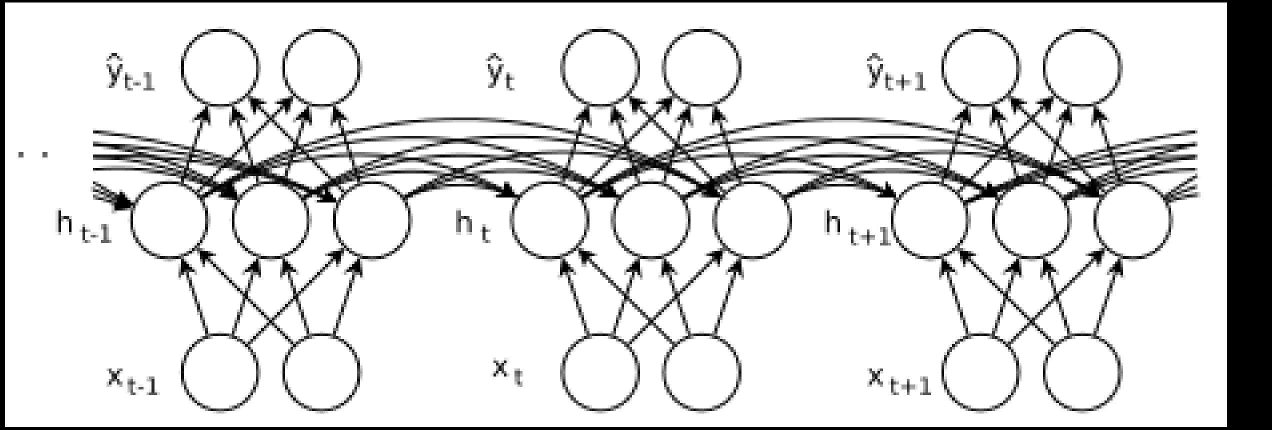
Advantages：

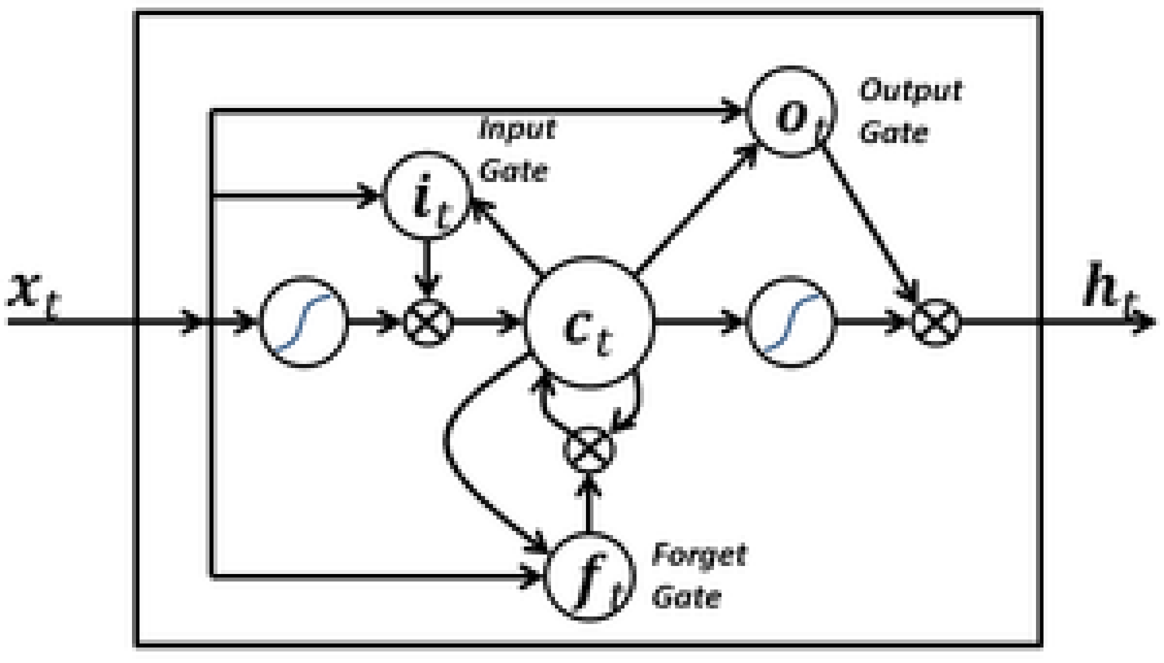
1. Less data to process；
2. High speed；
3. Work in Low-light environment；

Temporal Dependency Between Video Frames

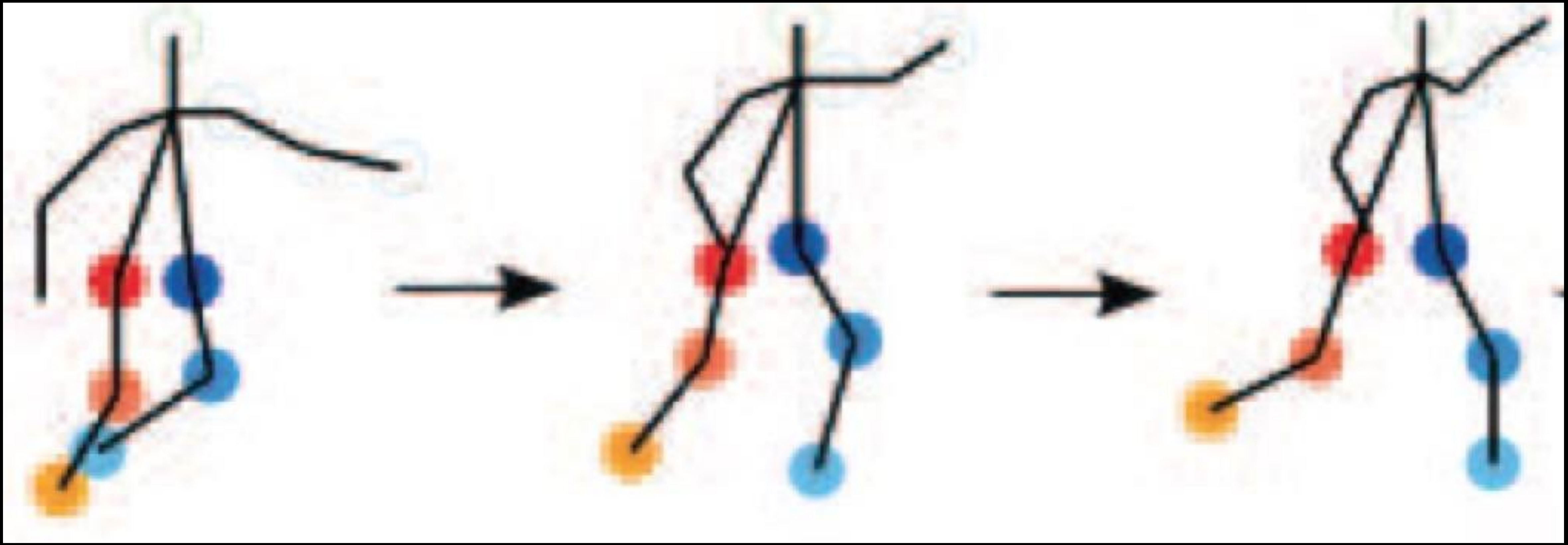


RNN and LSTM for Contextual Modelling



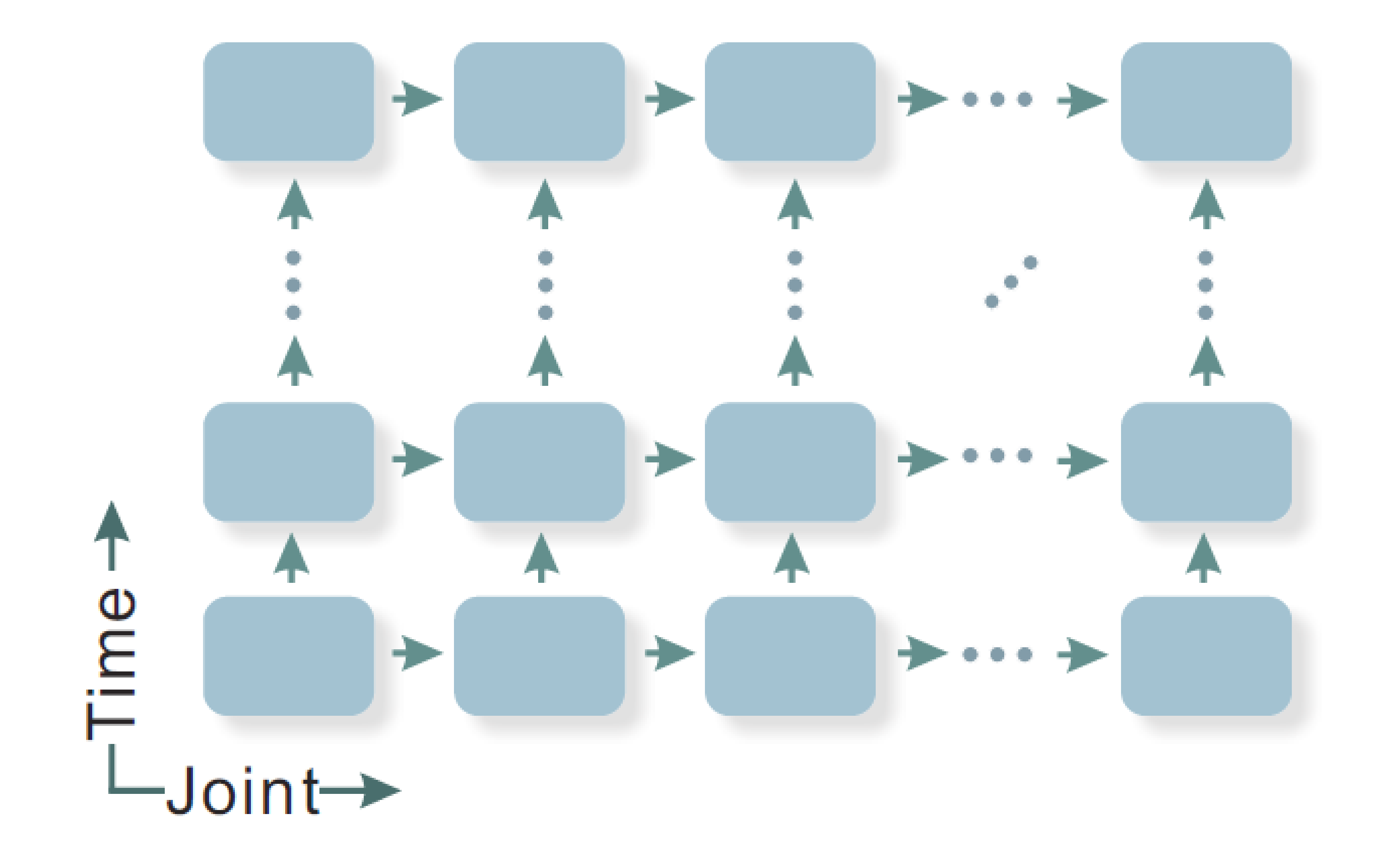


Contextual Dependency Between Different Joints



Wang et al. CVPR 2013

Spatial-temporal LSTM



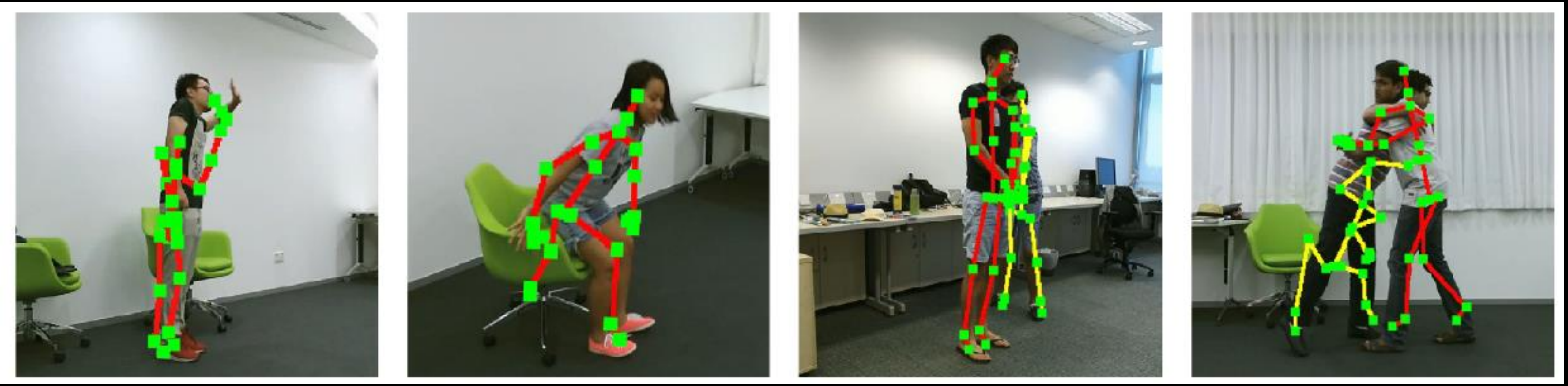
Formula

$$\begin{pmatrix} i_{j,t} \\ f^S_{j,t} \\ f^T_{j,t} \\ o_{j,t} \\ u_{j,t} \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} \left(M \begin{pmatrix} x_{j,t} \\ h_{j-1,t} \\ h_{j,t-1} \end{pmatrix} \right)$$

$$c_{j,t} = i_{j,t} \odot u_{j,t} + f^S_{j,t} \odot c_{j-1,t} + f^T_{j,t} \odot c_{j,t-1}$$

$$h_{j,t} = o_{j,t} \odot \tanh(c_{j,t})$$

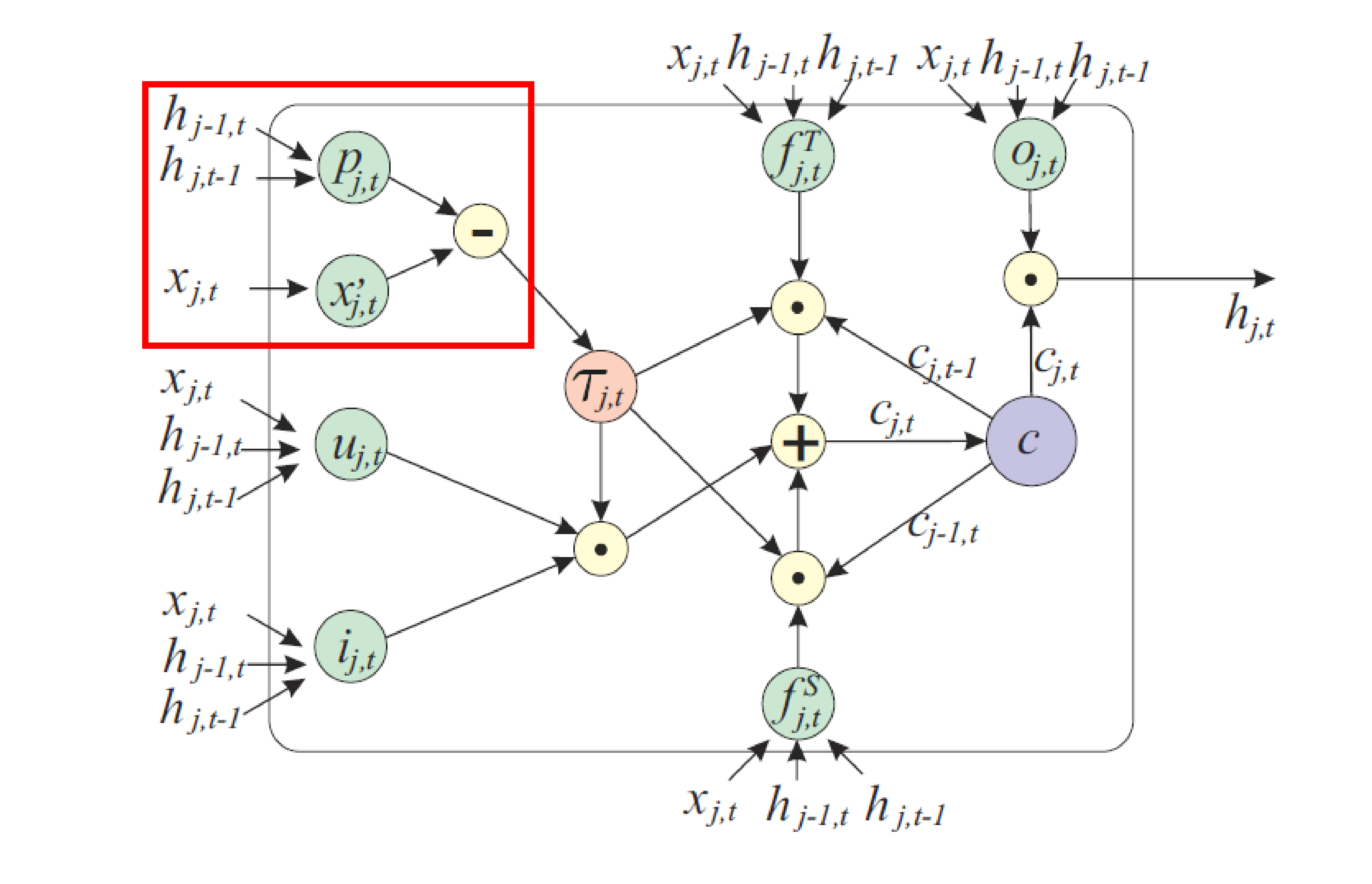
Joint Locations Can Be Noisy



Robust Trust Gates

- Noise and occlusion may contaminate the model and limit the performance.
- Add a new trust gate to the LSTM unit which analyzes the reliability of the input at each spatio-temporal step.
- Such trust gates can adaptively block the unreliable inputs and prevent the memory cell from updating.

Robust Trust Gate



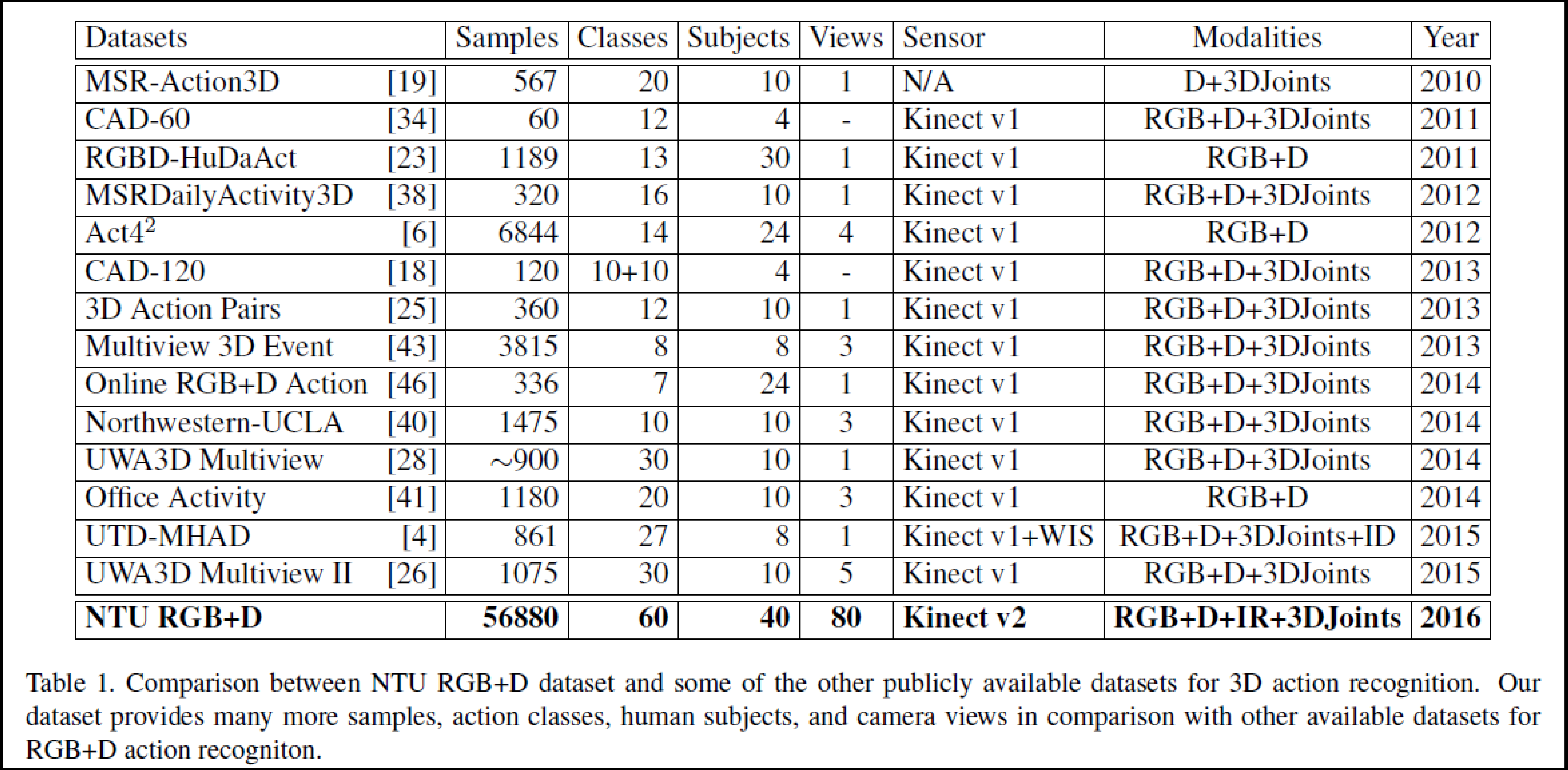
Formulation

$$p_{j,t} = \tanh\left(M_p \begin{pmatrix} h_{j-1,t} \\ h_{j,t-1} \end{pmatrix} \right)$$

$$x'_{j,t} = \tanh\left(M_x\left(x_{j,t}\right)\right)$$
$$\tau_{j,t} = G(x'_{j,t} - p_{j,t})$$

$$G(z) = \exp(-\lambda z^2)$$

NTU RGB-D Dataset



Datasets		Samples	Classes	Subjects	Views	Sensor	Modalities	Year
MSR-Action3D	[19]	567	20	10	1	N/A	D+3DJoints	2010
CAD-60	[34]	60	12	4	-	Kinect v1	RGB+D+3DJoints	2011
RGBD-HuDaAct	[23]	1189	13	30	1	Kinect v1	RGB+D	2011
MSRDailyActivity3D	[38]	320	16	10	1	Kinect v1	RGB+D+3DJoints	2012
Act4^2	[6]	6844	14	24	4	Kinect v1	RGB+D	2012
CAD-120	[18]	120	10+10	4	-	Kinect v1	RGB+D+3DJoints	2013
3D Action Pairs	[25]	360	12	10	1	Kinect v1	RGB+D+3DJoints	2013
Multiview 3D Event	[43]	3815	8	8	3	Kinect v1	RGB+D+3DJoints	2013
Online RGB+D Action	[46]	336	7	24	1	Kinect v1	RGB+D+3DJoints	2014
Northwestern-UCLA	[40]	1475	10	10	3	Kinect v1	RGB+D+3DJoints	2014
UWA3D Multiview	[28]	∼900	30	10	1	Kinect v1	RGB+D+3DJoints	2014
Office Activity	[41]	1180	20	10	3	Kinect v1	RGB+D	2014
UTD-MHAD	[4]	861	27	8	1	Kinect v1+WIS	RGB+D+3DJoints+ID	2015
UWA3D Multiview II	[26]	1075	30	10	5	Kinect v1	RGB+D+3DJoints	2015
NTU RGB+D		**56880**	**60**	**40**	**80**	**Kinect v2**	**RGB+D+IR+3DJoints**	**2016**

Table 1. Comparison between NTU RGB+D dataset and some of the other publicly available datasets for 3D action recognition. Our dataset provides many more samples, action classes, human subjects, and camera views in comparison with other available datasets for RGB+D action recogniton.

Results

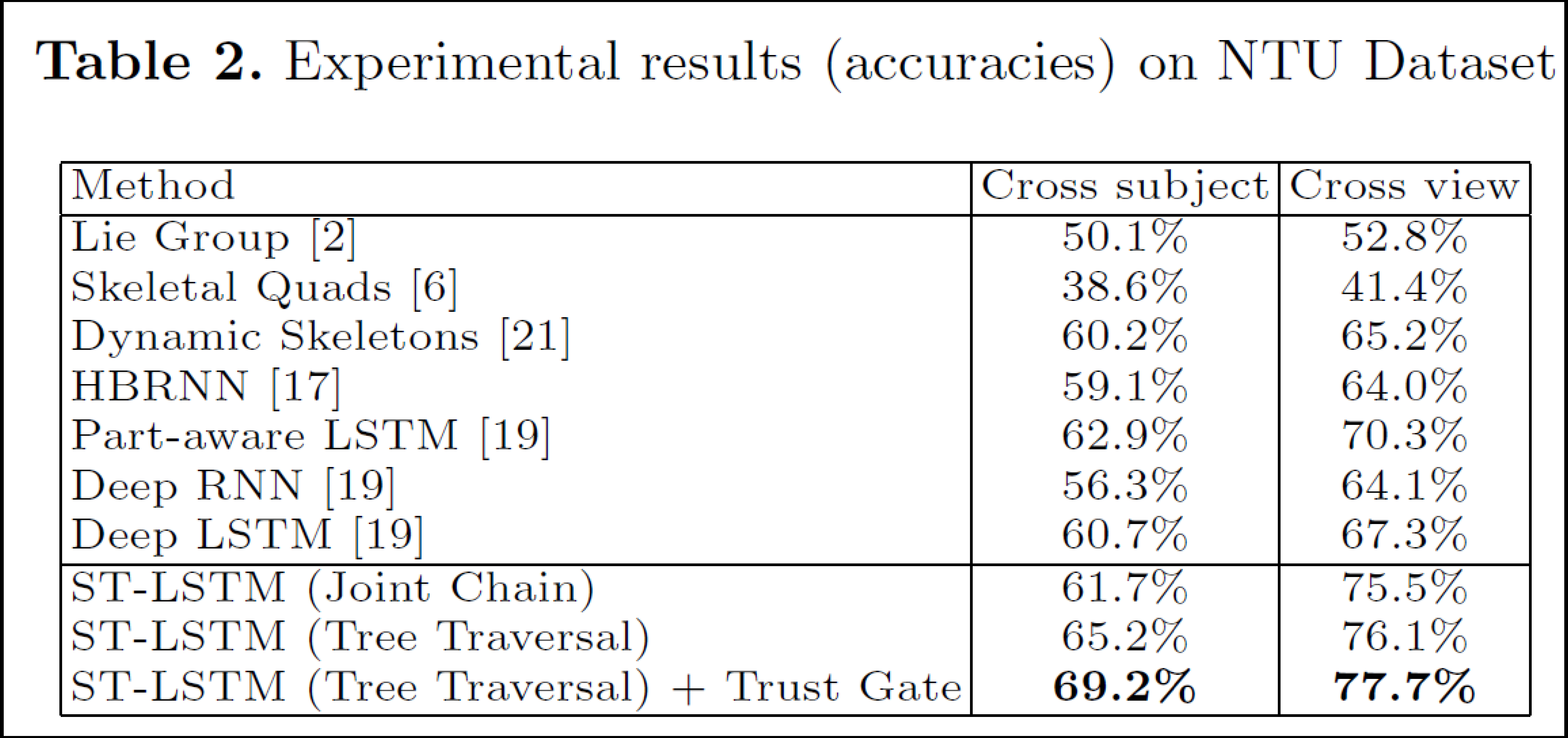


Table 2. Experimental results (accuracies) on NTU Dataset

Method	Cross subject	Cross view
Lie Group [2]	50.1%	52.8%
Skeletal Quads [6]	38.6%	41.4%
Dynamic Skeletons [21]	60.2%	65.2%
HBRNN [17]	59.1%	64.0%
Part-aware LSTM [19]	62.9%	70.3%
Deep RNN [19]	56.3%	64.1%
Deep LSTM [19]	60.7%	67.3%
ST-LSTM (Joint Chain)	61.7%	75.5%
ST-LSTM (Tree Traversal)	65.2%	76.1%
ST-LSTM (Tree Traversal) + Trust Gate	**69.2%**	**77.7%**

Conclusions

- Designing neural networks with gates to adpatively extract information in the inference process is a promising research direction.
- Our works have demonstrated the effectiveness on several important tasks.
- Future works can be done to design principles with more general design mechanism.