Human Activity Recognition with Pose-driven Attention to RGB

BMVC 2018 Newcastle upon Tyne

Fabien Baradel[1], Christian Wolf[1,2], Julien Mille[3]

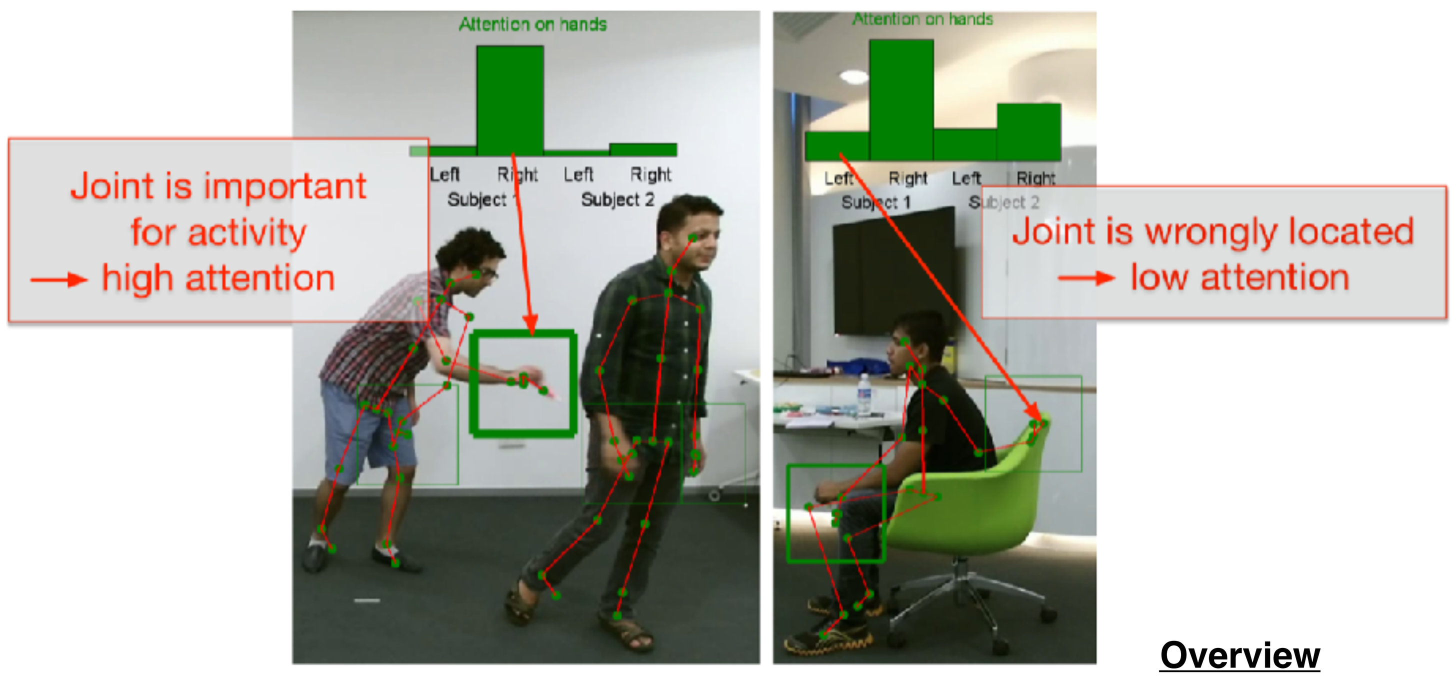
1 Univ. Lyon, INSA-Lyon, CNRS, LIRIS, F-69621, Villeurbanne, France.
2 INRIA, CITI Laboratory, Villeurbanne, France.
3 Laboratoire d'Informatique de l'Univ. de Tours, INSA Centre Val de Loire, 41034 Blois, France.

fabien.baradel@liris.cnrs.fr

PROBLEM DEFINITION & MOTIVATIONS

Problem statement:

How can an **attention mechanism select the most discriminative parts of the video**?



Main challenges

- **High dimensional** data
- **Spatio-Temporal** information
- **Noise** in the human pose

Overview

- **Video** Understanding
- **Human** Action Recognition
- Captured by **Microsoft Kinect3D** (*3D human pose - RGB - Depth*)

MAIN IDEA

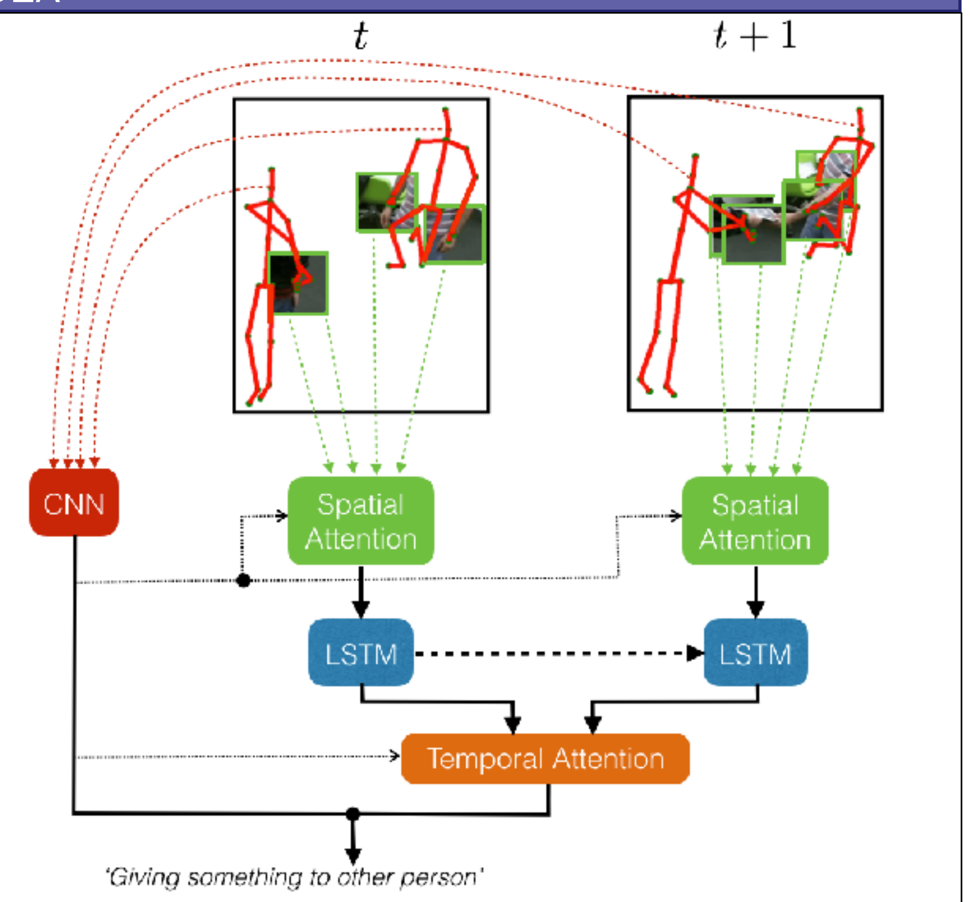
- Two modalities
  - ✓ *3D skeleton coordinates*
  - ✓ *RGB frames*
- **Two stream model**

Pose

- **Convolution** features over *space-time*
- **Long range dependencies** are modeled with multiple layers of abstraction *instead of a flat hidden RNN state*

RGB

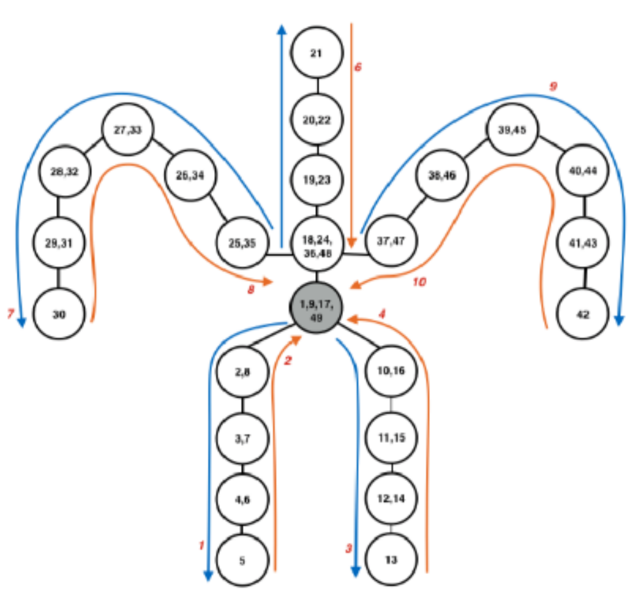
- **Spatial attention mechanism** over RGB hands crops
- Spatial attention adjusted at each timestep
- **Conditioned on pose features**
- **Temporal Attention** on LSTM features



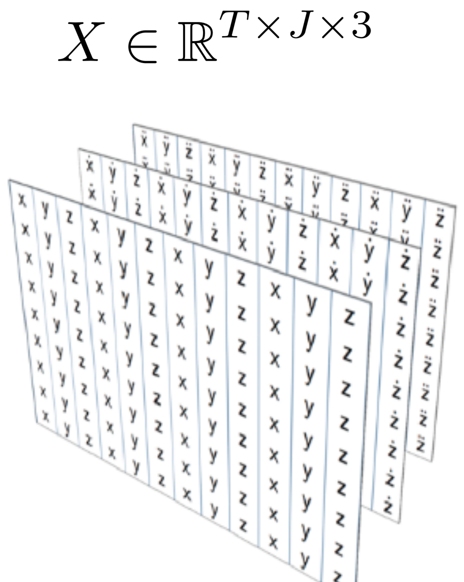
TWO STREAM MODEL

CONVNET ON SKELETON

Topological ordering



$X \in \mathbb{R}^{T \times J \times 3}$

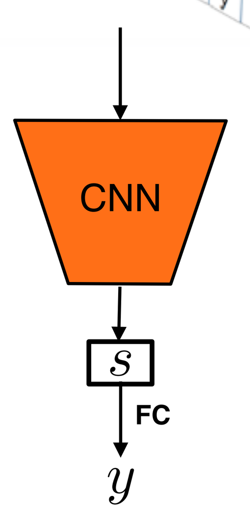


Preprocessing

- The matrix X is filled with a **topological (neighborhood preserving) ordering**
- Position, Acceleration and Velocity

Model

- **Convolutions** extract *spatio-temporal features*
- S : **pose features** (~body motion)



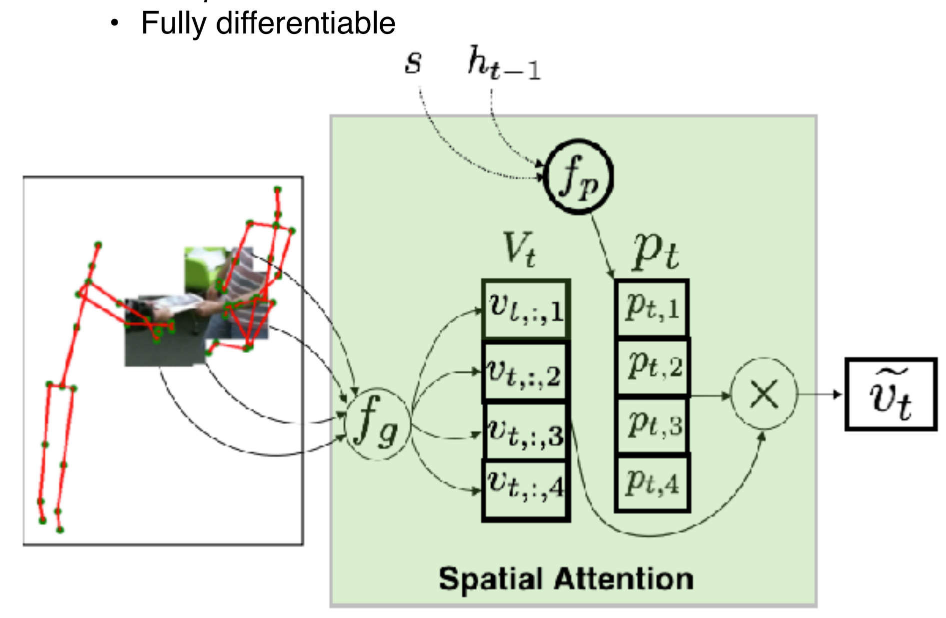
$T =$ # timestep

$J =$ (# subjects) x (# joints after repeated topological ordering) x (# coordinates dimension)

ATTENTION ON HANDS

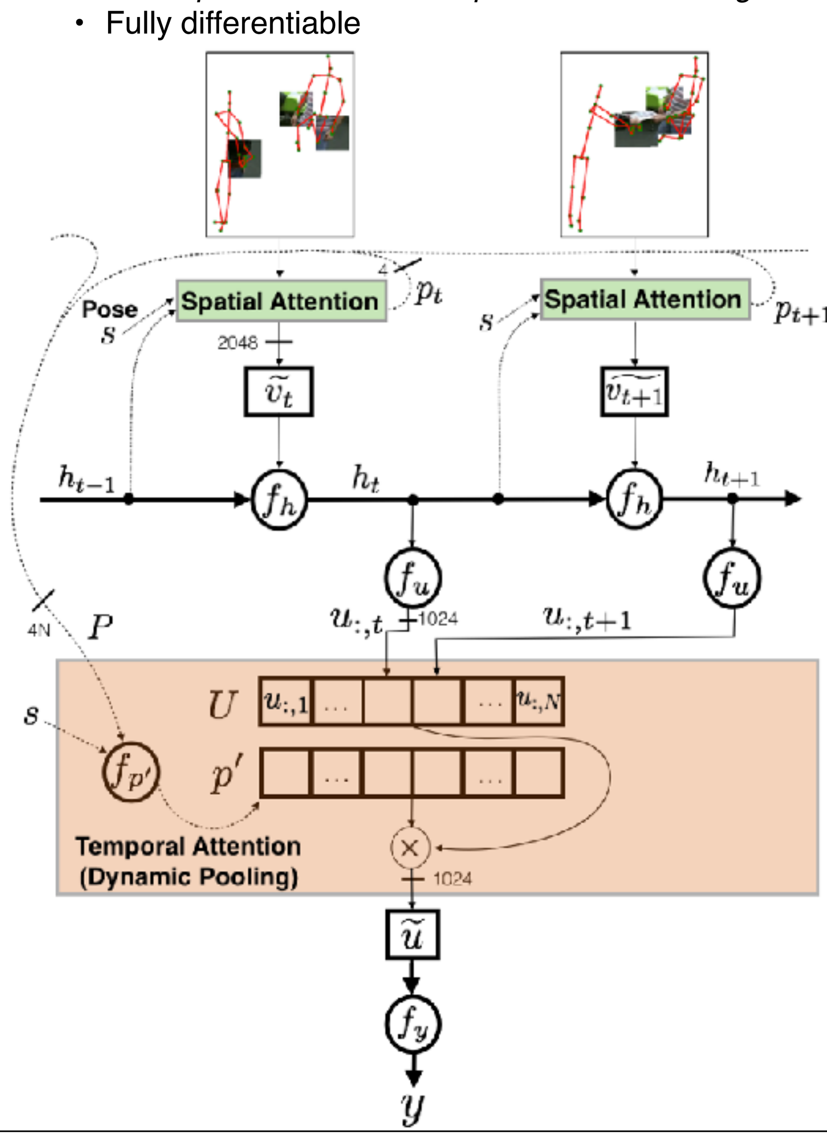
Spatial Attention around Hands crops

- *Inception features* from **RGB crops around hands**
- **Attention weights** computed given
  - ✓ *previous hidden state of RNN* h_{t-1}
  - ✓ *pose features* S
- Fully differentiable



Temporal Attention on LSTM features

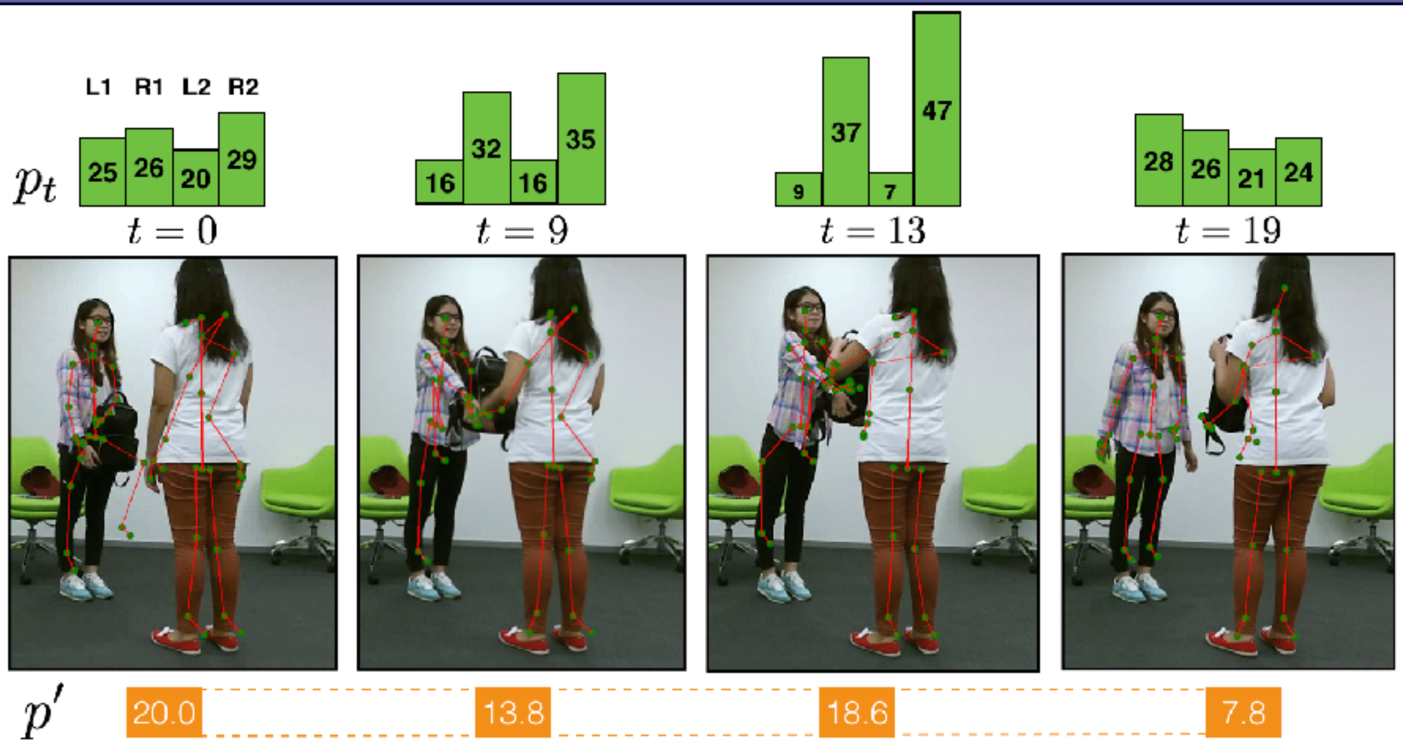
- Can be seen as a **dynamic pooling**
- **Weighted average** of features over all frames
- Given *pose features* and *spatial attention weights*
- Fully differentiable

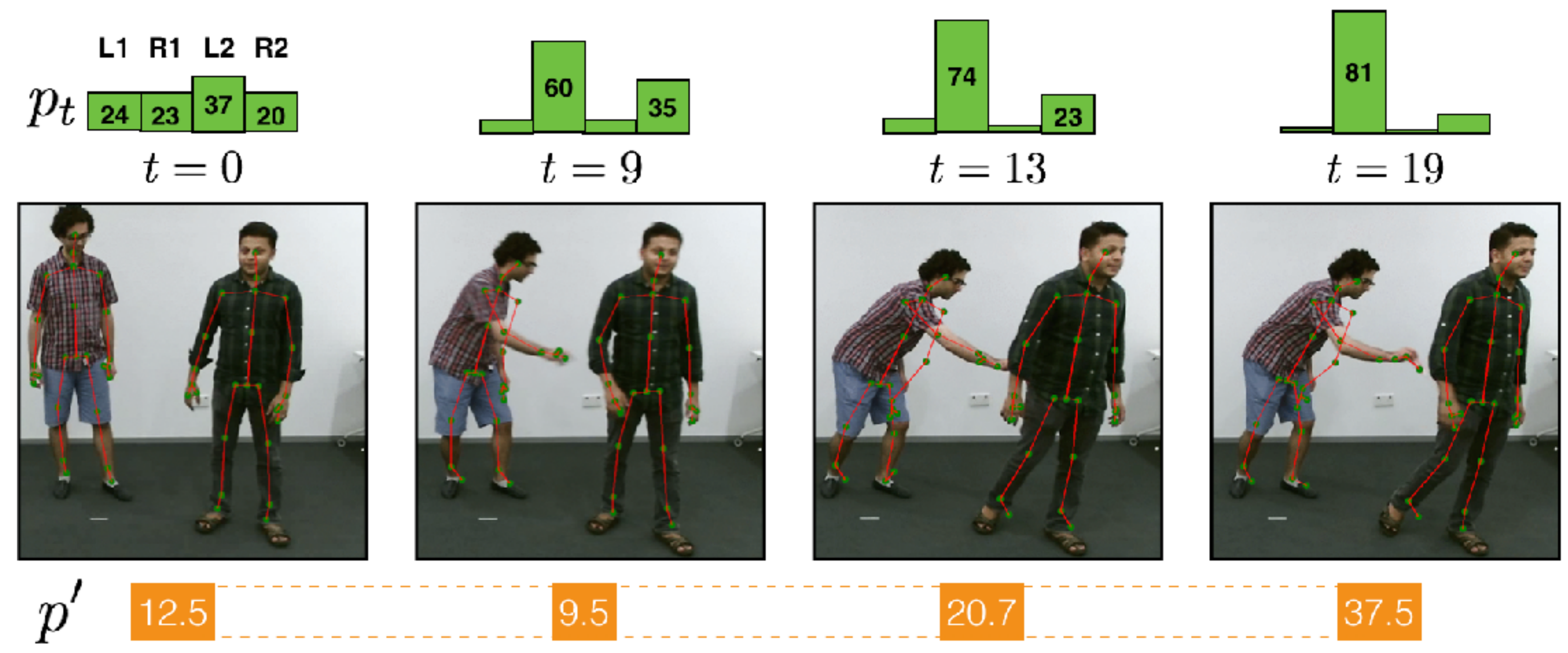


Glossary:

f_g: Inception feature vector
p_t: Spatial Attention weights for each hand
$\tilde{v}_t$: Output of the Spatial Attention framework - Input of the LSTM
f_h: LSTM
f_u: MLP
p': Temporal Attention weights for each LSTM features
f_y: Classifier
$u_{:,t}$: LSTM features

VISUALIZATION OF THE ATTENTION PROCESS





EXPERIMENTAL RESULTS

Comparison

- **State of the art on NTU RGB+D** (NTU) (~57'000 videos - 60 classes)
- First to combine **3D skeleton data and RGB frames** on NTU
- Representations learned on NTU are transferable
  - ✓ **Transfer learning on smaller datasets**
  - ✓ State of the art on SBU Kinetics Interaction
  - ✓ Close to state of the art on MSR Daily Activity

Ablation Study

- **Topological ordering matters** for skeleton data
- **Attention mechanism has a high impact** on RGB only stream
  - ✓ Spatial Attention : + ~ 4 points
  - ✓ Spatio-Temporal Attention : + ~ 13 points
- Still a significant impact on the **two stream model**
  - ✓ Spatial Attention : + ~ 1.8 points
  - ✓ **Spatio-Temporal Attention : + ~ 2.5 points**

Effect of joint ordering on NTU

Methods	CS	CV	Avg
Random joint order	75.5	83.2	79.4
Topological order w/o double entries	76.2	83.9	80.0
Topological order	77.1	84.5	80.8

NTU-RGB+D

Methods	Pose	RGB	CS	CV	Avg
Lie Group	✓	-	50.1	52.8	51.5
Skeleton Quads	✓	-	38.6	41.4	40.0
Dynamic Skeletons	✓	-	60.2	65.2	62.7
HBRNN	✓	-	59.1	64.0	61.6
Deep LSTM	✓	-	60.7	67.3	64.0
Part-aware LSTM	✓	-	62.9	70.3	66.6
ST-LSTM + TrustG.	✓	-	69.2	77.7	73.5
STA-LSTM	✓	-	73.2	81.2	77.2
Ensemble LSTM	✓	-	74.6	81.3	78.0
GCA-LSTM	✓	-	74.4	82.8	78.6
JTM	✓	-	76.3	81.1	78.7
MTLN	✓	-	79.6	84.8	82.2
VA-LSTM	✓	-	79.4	87.6	83.5
View-invariant	✓	-	80.0	87.2	83.6
DSSCA-SSLM	✓	✓	74.9	-	-
STA-Hands	✓	✓	82.5	88.6	85.6
Hands Attention	✓	✓	84.8	90.6	87.7
C3D	-	✓	63.5	70.3	66.9
Resnet50+LSTM	-	✓	71.3	80.2	75.8
Ours (pose only)	✓	-	**77.1**	**84.5**	**80.8**
Ours (RGB only)	-	✓	**75.6**	**80.5**	**78.1**
Ours (pose + RGB)	✓	✓	**84.8**	**90.6**	**87.7**

SBU Kinect Interaction

Methods	Pose	RGB	Depth	Acc.
Raw skeleton	✓	-	-	79.4
Joint feature	✓	-	-	86.9
Co-occurence RNN	✓	-	-	90.4
STA-LSTM	✓	-	-	91.5
ST-LSTM + TrustG.	✓	-	-	93.3
DSPM	✓	✓	✓	93.4
VA-LSTM	✓	-	✓	97.5
Ours (pose only)	✓	-	-	**90.5**
Ours (RGB only)	-	✓	✓	**72.0**
Ours (pose + RGB)	✓	✓	✓	**94.1**

MSR Daily Interaction

Methods	Pose	RGB	Depth	Acc.
Action Ensemble	✓	-	-	68.0
Efficient Pose-Based	✓	-	-	73.1
Moving Pose	✓	-	-	73.8
Moving Poselets	✓	-	-	74.5
MP	✓	-	-	79.9
Depth Fusion	-	-	✓	88.8
MMMP	✓	-	✓	91.3
DL-GSGC	✓	-	✓	95.0
DSSCA-SSLM	-	✓	✓	97.5
Ours (pose only - no FT)	✓	-	-	**72.2**
Ours (pose only)	✓	-	-	**74.6**
Ours (RGB only)	-	✓	✓	**75.3**
Ours (pose + RGB)	✓	✓	✓	**90.0**

Ablation study on NTU

	Methods	Pose	RGB	Attention			CS	CV	Avg
				Spatial	Temporal	Pose			
A	Pose only	✓	-	-	-	-	77.1	84.5	80.8
B	RGB only, no attention (sum of features)	-	✓	-	-	-	61.5	65.9	63.7
C	RGB only, no attention (concat of features)	-	✓	-	-	-	63.2	67.2	65.2
E	RGB only + spatial attention	o	✓	✓	-	✓	67.4	71.2	69.3
G	RGB only + spatial-temporal attention	o	✓	✓	✓	✓	75.6	80.5	78.1
H	Multi-modal, no attention (A+B)	✓	✓	-	-	-	83.0	88.5	85.3
I	Multi-modal, spatial attention (A+E)	✓	✓	✓	-	✓	84.1	90.0	87.1
K	Multi-modal, spatial-temporal attention (A+E)	✓	✓	✓	✓	✓	84.8	90.6	87.7