ICPR 2020
25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION
Milan, Italy 10 | 15 January 2021

25th International Conference on Pattern Recognition, Milan, Italy, Jan.10-15, 2021

Attention-Oriented Action Recognition for Real-Time Human-Robot Interaction

Ziyang Song[1], Ziyi Yin[1], Zejian Yuan[1], Chong Zhang[2], Wanchao Chi[2], Yonggen Ling[2], Shenghao Zhang[2]

[1]Xi'an Jiaotong University, Xi'an, China

[2]Tencent Robotics X, Shenzhen, China

西安交通大學
XI'AN JIAOTONG UNIVERSITY

Tencent 腾讯

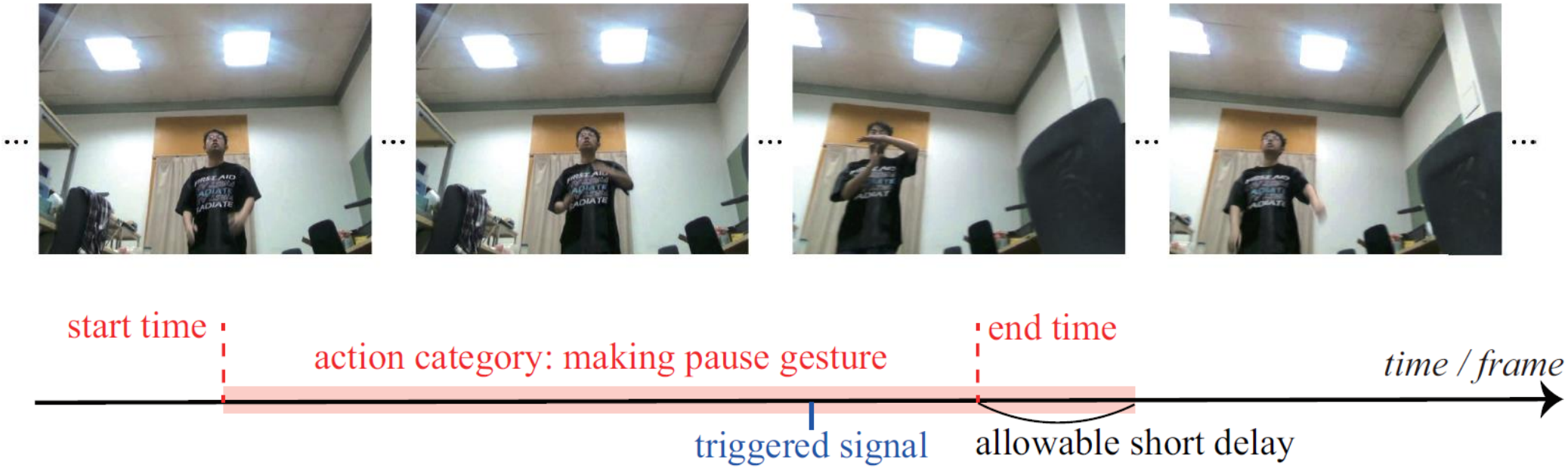
Background

Action Recognition in HRI

- Public service
- Home service
- Entertainment

Task

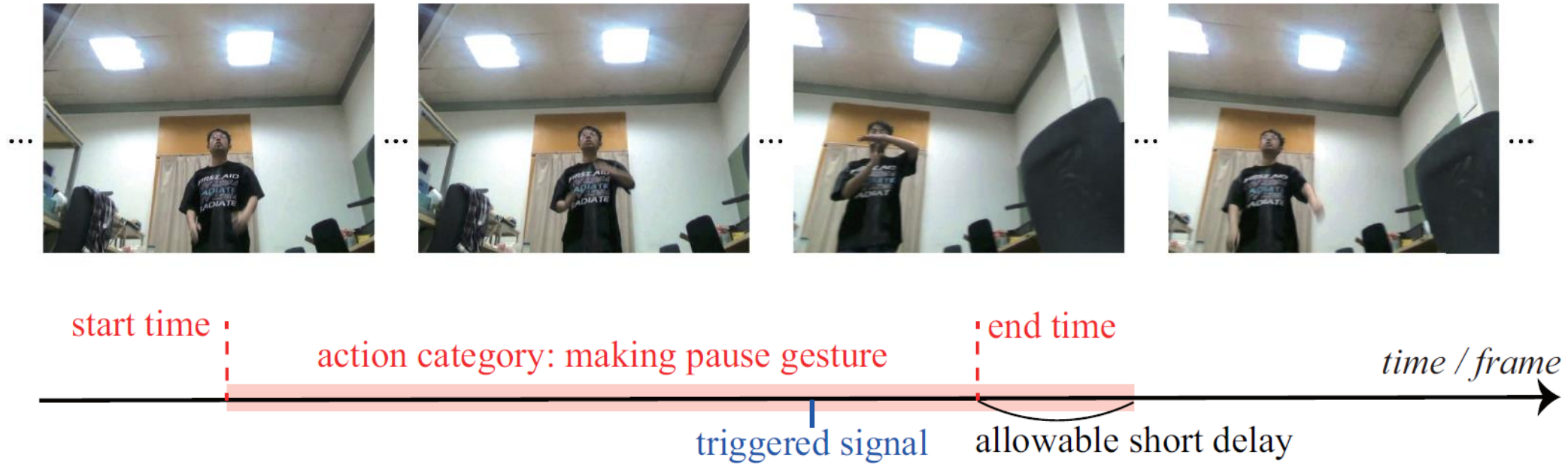
Recognize the interactor's actions from video stream



Task

Recognize the interactor’s actions from video stream

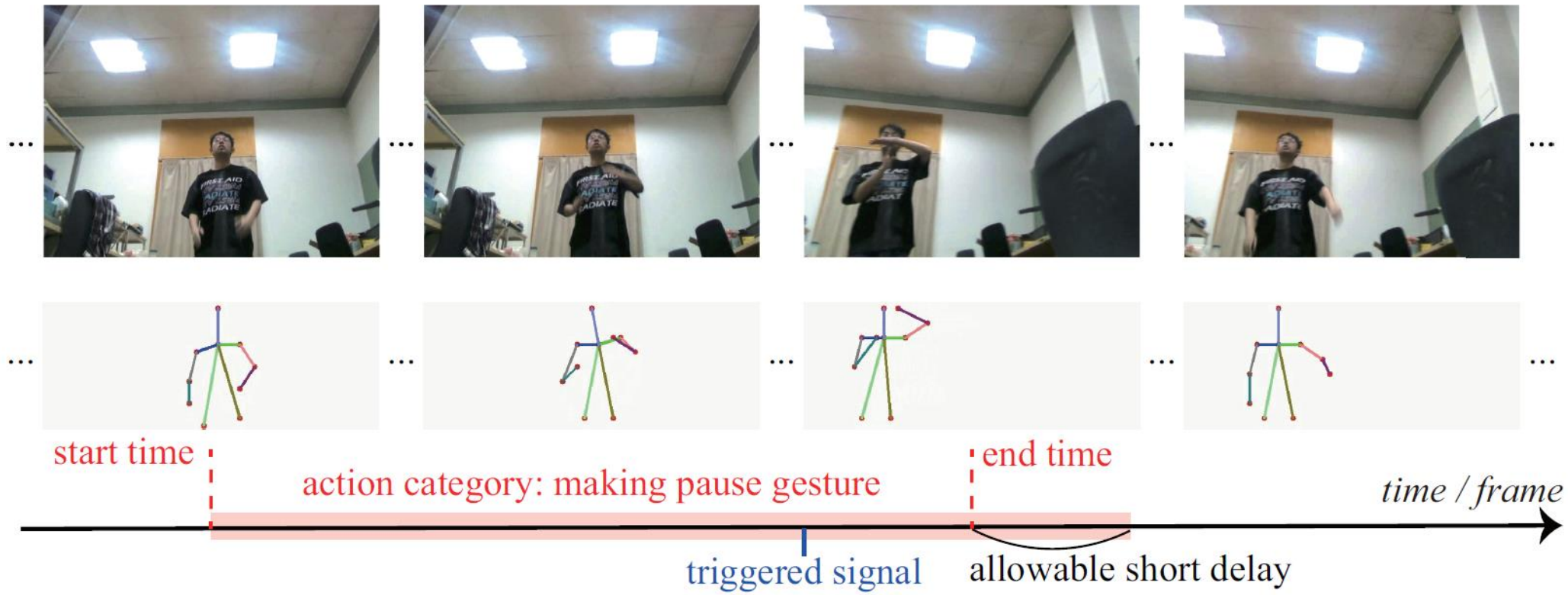
- Specific scenes and camera viewpoints
- Real-time recognition on the mobile robot platform



Task

Two-step pipeline

- Pose estimation
- Skeleton-based action recognition



Current Work and Limitations

Pose Estimation

- Single-person method [1]
  - Incapability in multi-person scene
- Multi-person method
  - Bottom-up [2]
    - Resource consumption on irrelevant people
  - Top-down [3]
    - Selection module tough to design

[1] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, “Convolutional pose machines,” in CVPR, 2016.
[2] Z. Cao, G. Martinez, T. Simon, S.-E. Wei, and Y. Sheikh, “Openpose: Realtime multi-person 2d pose estimation using part affinity fields,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019.
[3] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, “Cascaded pyramid network for multi-person pose estimation,” in CVPR, 2018.

Current Work and Limitations

Pose Estimation

- Single-person method
  - Incapability in multi-person scene
- Multi-person method
  - Bottom-up
    - Resource consumption on irrelevant people
  - Top-down
    - Selection module tough to design

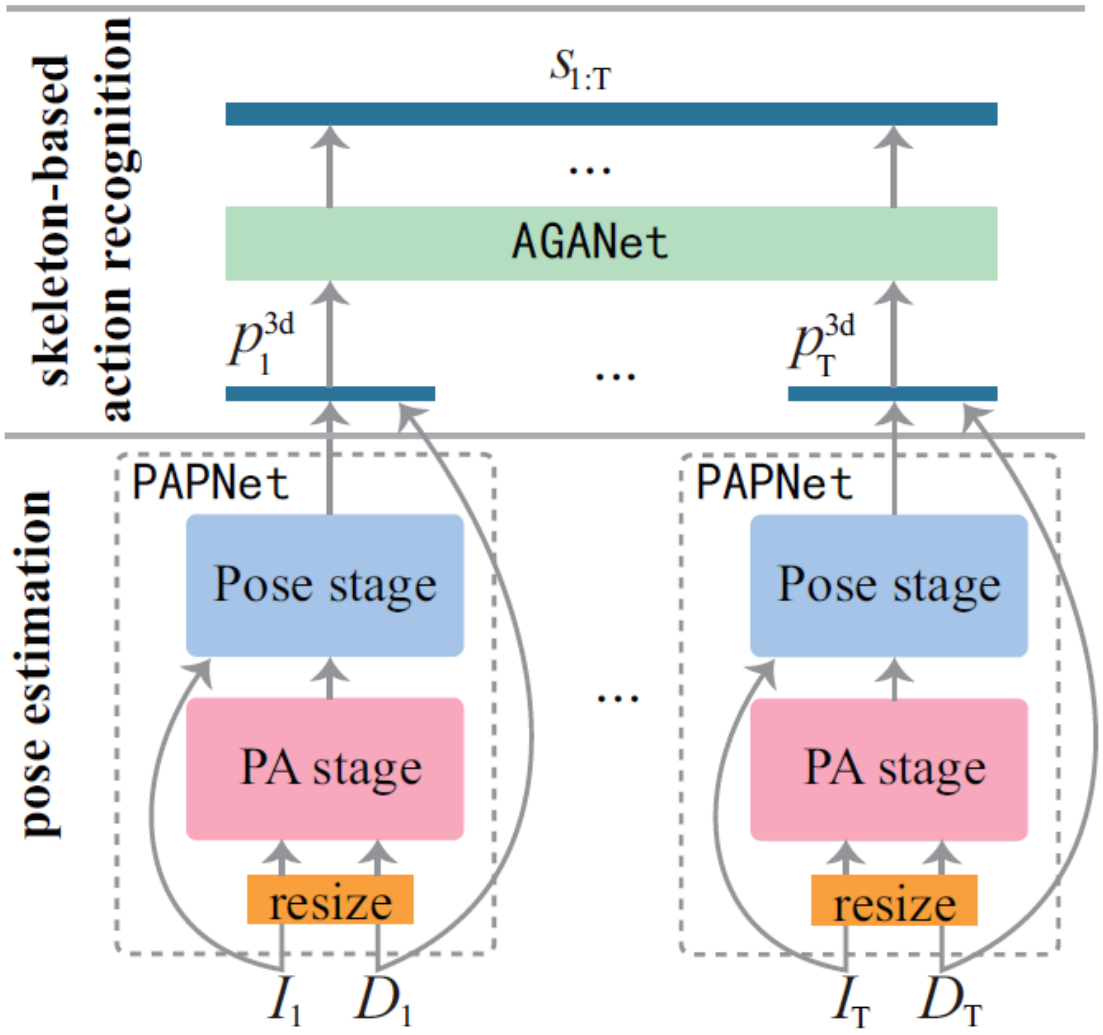
Skeleton-Based Action Recognition [1][2][3][4]

- Lack of explainability

[1] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3d action recognition," in CVPR, 2017.
[2] Y. Li, C. Lan, J. Xing, W. Zeng, C. Yuan, and J. Liu, "Online human action detection using joint classification-regression recurrent neural networks," in ECCV, 2016.
[3] H. Wang and L. Wang, "Beyond joints: Learning representations from primitive geometries for skeleton-based action recognition and detection," IEEE Transactions on Image Processing, vol. 27, pp. 4382–4394, 2018.
[4] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in AAAI, 2018.

Proposed Method

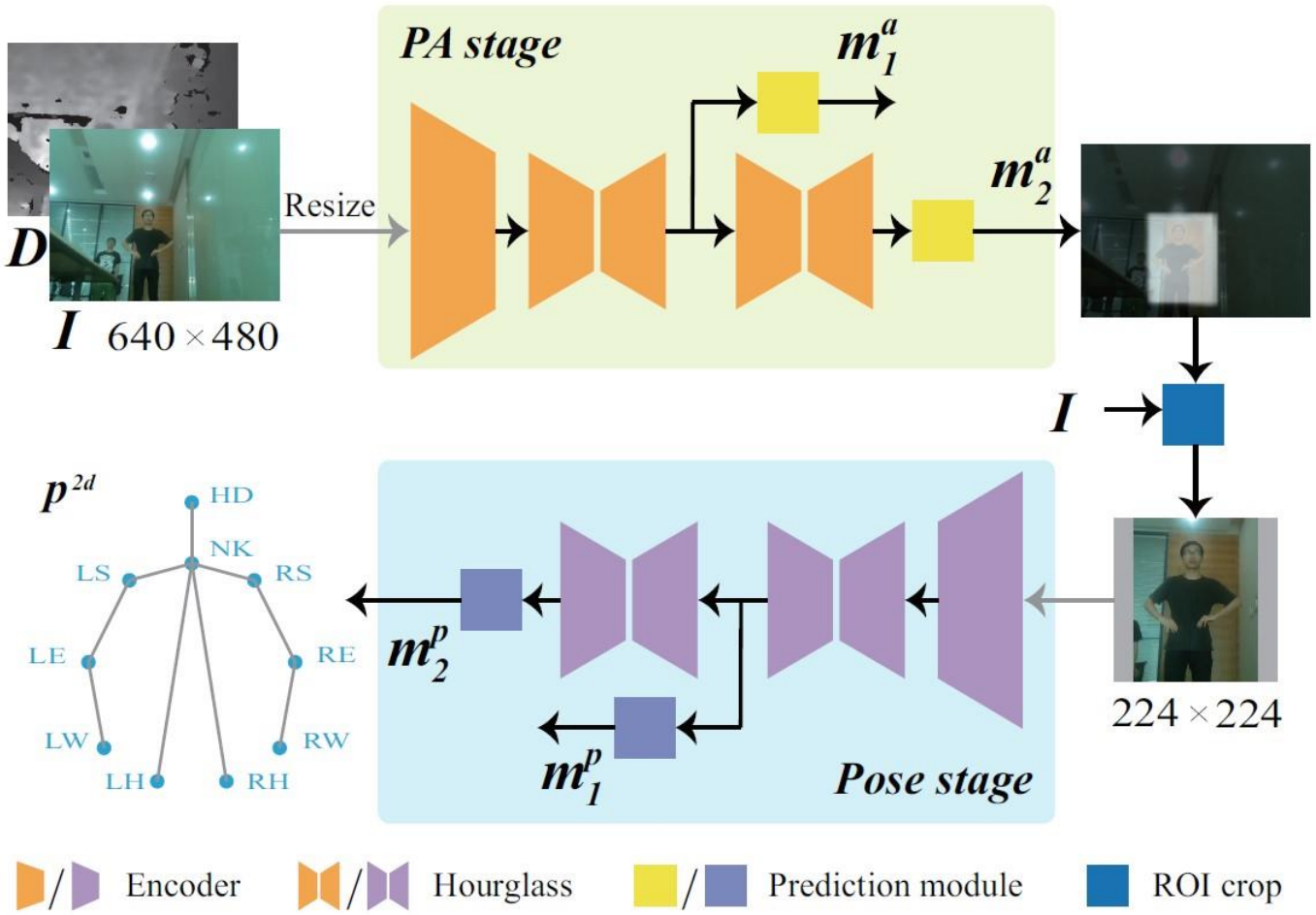
Attention-oriented multi-level network framework



Proposed Method

Pre-Attention Pose Network (PAPNet)

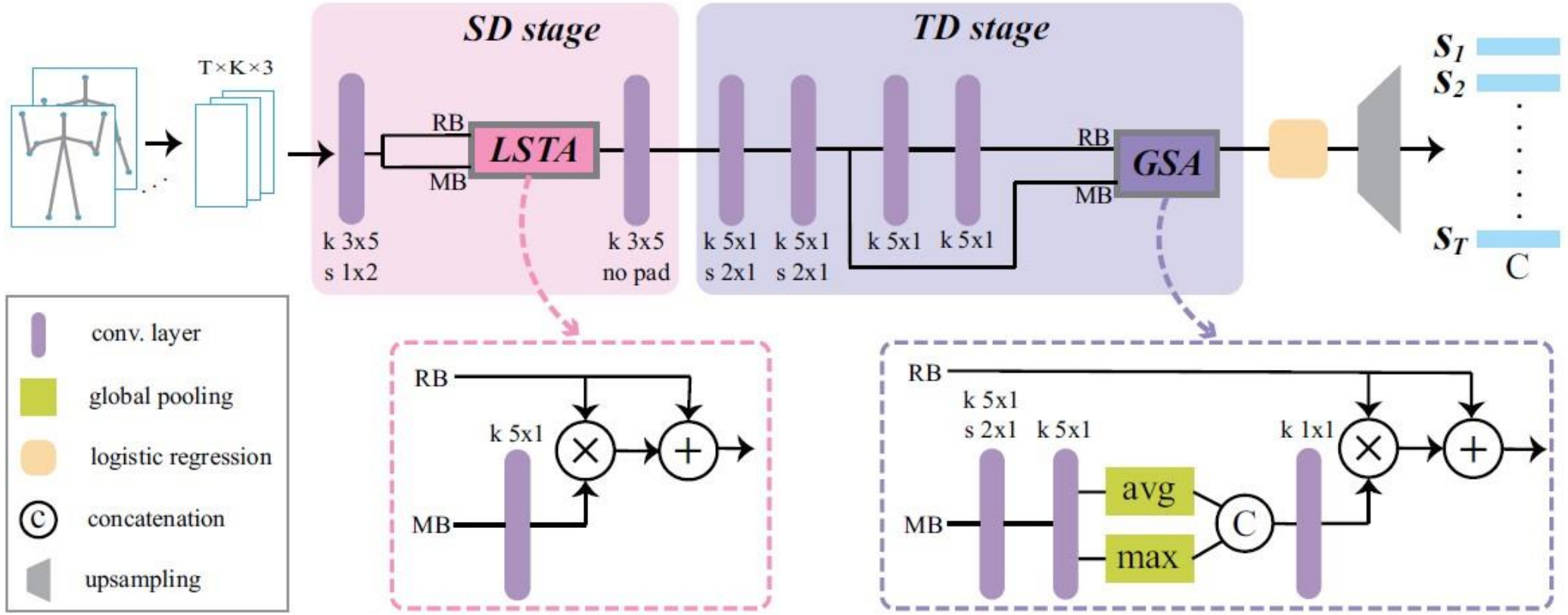
- Learning to roughly focusing on the interactor with Pre-Attention



Proposed Method

Attention-Guided Action Network (AGANet)

- Local Spatial-Temporal Attention (LSTA)
- Global Semantic Attention (GSA)



Dataset

Action-in-Interaction Dataset (AID)

- RGB + Depth
- Simulating mobile robot's viewpoints
- 10 categories
- 20 subjects
- 1031 action instances

1	Raise left hand
2	Raise right hand
3	Swing left hand
4	Swing right hand
5	Push forward with left hand
6	Push forward with right hand
7	Circle with left hand
8	Circle with right hand
9	Make pose gesture
10	Cross hands

Experiments

Evaluation of PAPNet

- Best efficiency far ahead (8.3 × smaller and 2.4 × faster than the 2nd place)
- Competitive accuracy
- Higher resolution output

model	parameters	fps	PCK@0.15
CPN v1 [1]	46.0M	33	**97.31**
CPN v2 [1]	27.0M	47	96.56
OpenPose v1 [2]	42.0M	15	96.70
OpenPose v2 [2]	11.6M	43	95.73
PAPNet	**1.4M**	**112**	96.00

[1] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for multi-person pose estimation," in CVPR, 2018.
[2] Z. Cao, G. Martinez, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019.

Experiments

Pre-Attention effects

- Robustness to position and scale changes
- Eliminating irrelevant people’s interference

Experiments

Evaluation of AGANet

- Leading accuracy

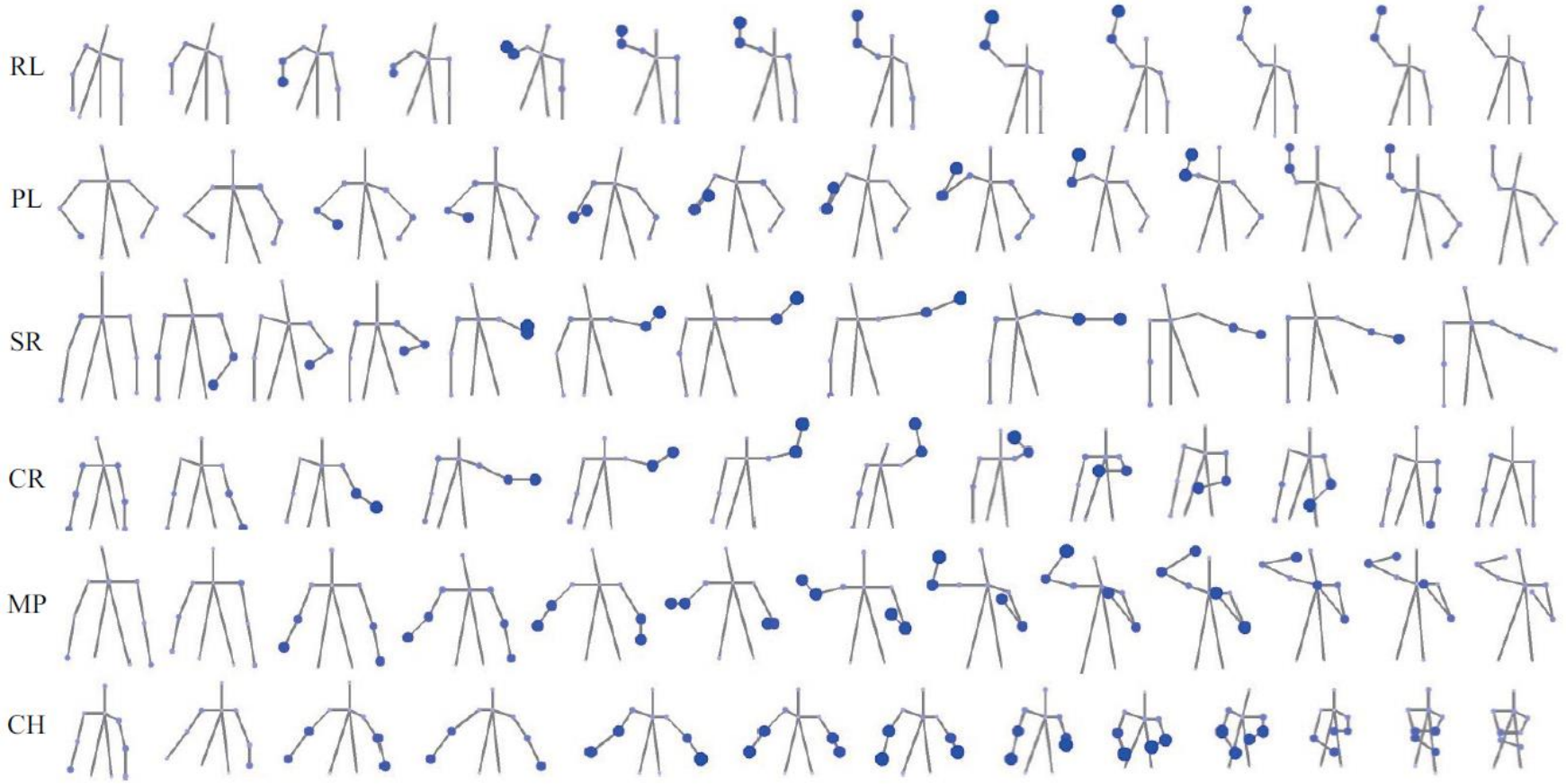
model	cAP	AP_{trig}	P_{trig}	R_{trig}
MTLN [1]	75.30	78.41	78.46	79.39
JCR-RNN [2]	66.95	73.79	79.08	73.60
Beyond joints [3]	76.86	82.28	83.07	83.07
ST-GCN [4]	81.63	87.56	87.72	87.41
base-AGANet	81.90	89.55	90.61	88.74
AGANet	**87.50**	**96.00**	**95.08**	**95.71**

[1] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3d action recognition," in CVPR, 2017.
[2] Y. Li, C. Lan, J. Xing, W. Zeng, C. Yuan, and J. Liu, "Online human action detection using joint classification-regression recurrent neural networks," in ECCV, 2016.
[3] H. Wang and L. Wang, "Beyond joints: Learning representations from primitive geometries for skeleton-based action recognition and detection," IEEE Transactions on Image Processing, vol. 27, pp. 4382–4394, 2018.
[4] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in AAAI, 2018.

Experiments

LSTA effects

- Focusing on main body parts involved in actions



Conclusion

Specifying a new action recognition task for HRI

Proposing an attention-oriented multi-level network framework for real-time action recognition

Construct a new dataset and define a new evaluation metric

For more details, please refer to:

*Z. Song, Z. Yin, Z. Yuan, C. Zhang, W. Chi, Y. Ling, S. Zhang. Attention-Oriented Action Recognition for Real-Time HRI. , 25th International Conference on Pattern Recognition (**ICPR**), Milan, Jan.10-15, **2021**.*