Progressive Teacher-student Learning for Early Action Prediction

Xionghui Wang, Jian-Fang Hu*, Jianhuang Lai,
Jianguo Zhang, and Wei-Shi Zheng

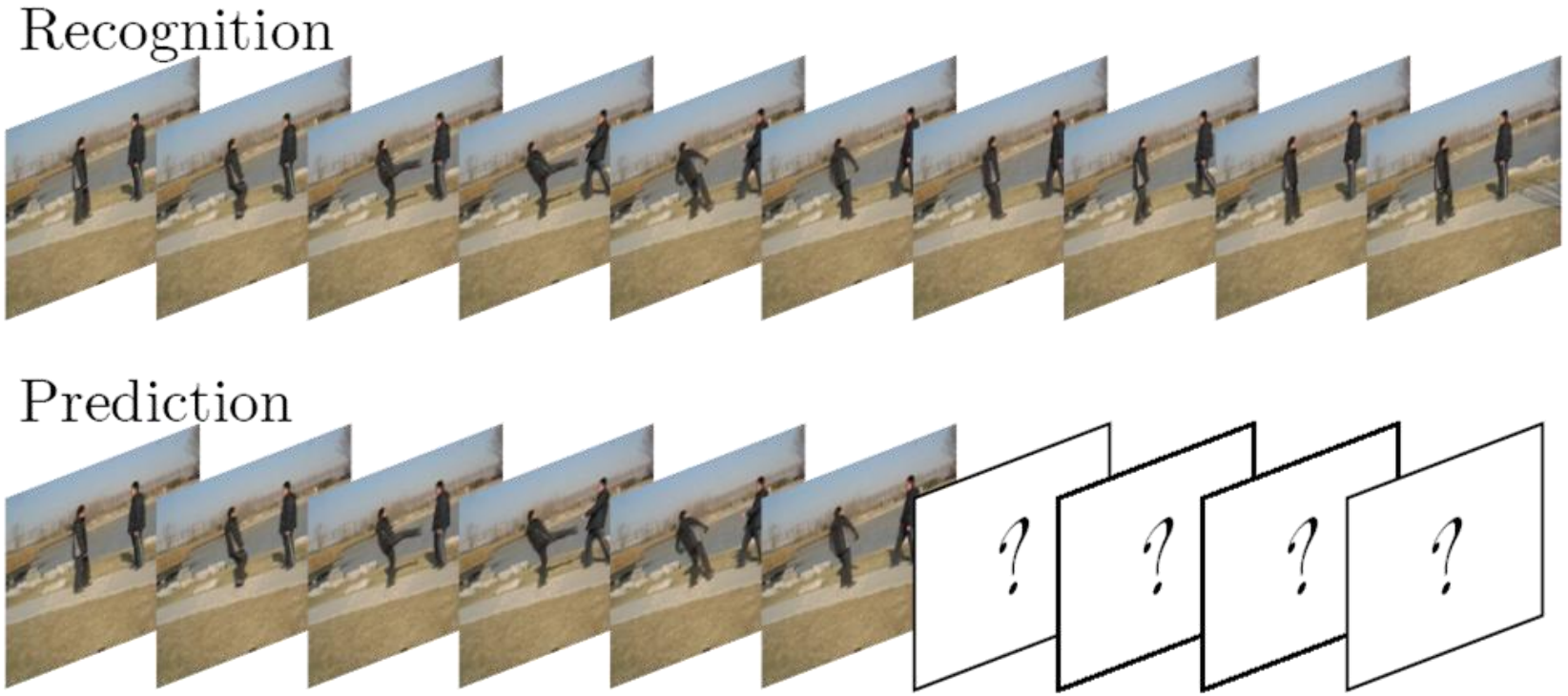
http://isee.sysu.edu.cn/~hujianfang/

SUN YAT-SEN University (中山大学)

OUTLINE

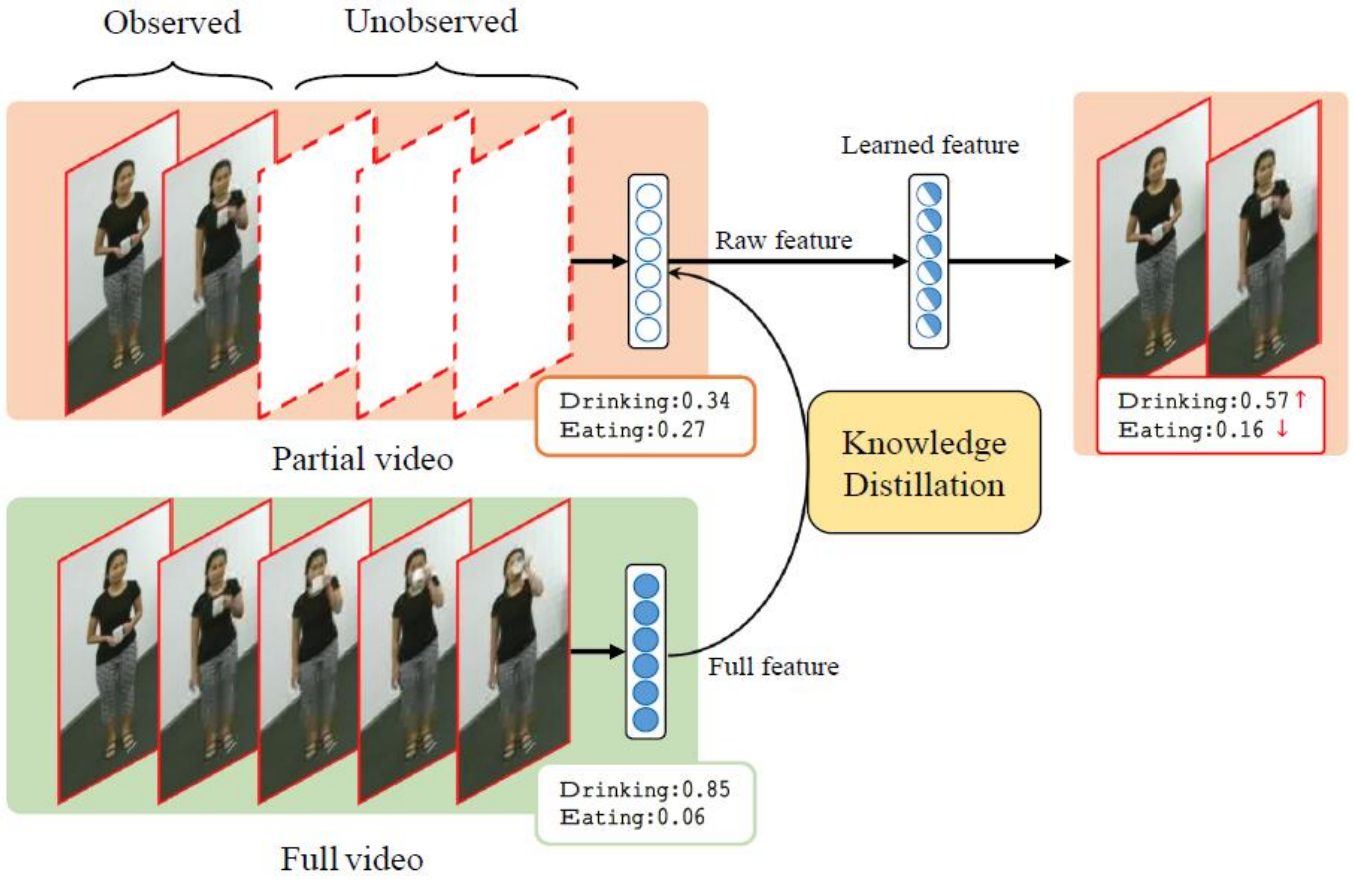
1. Introduction & Motivation
2. Feature Extraction
3. Teacher Student Learning
4. Summary

1. Introduction & Motivation



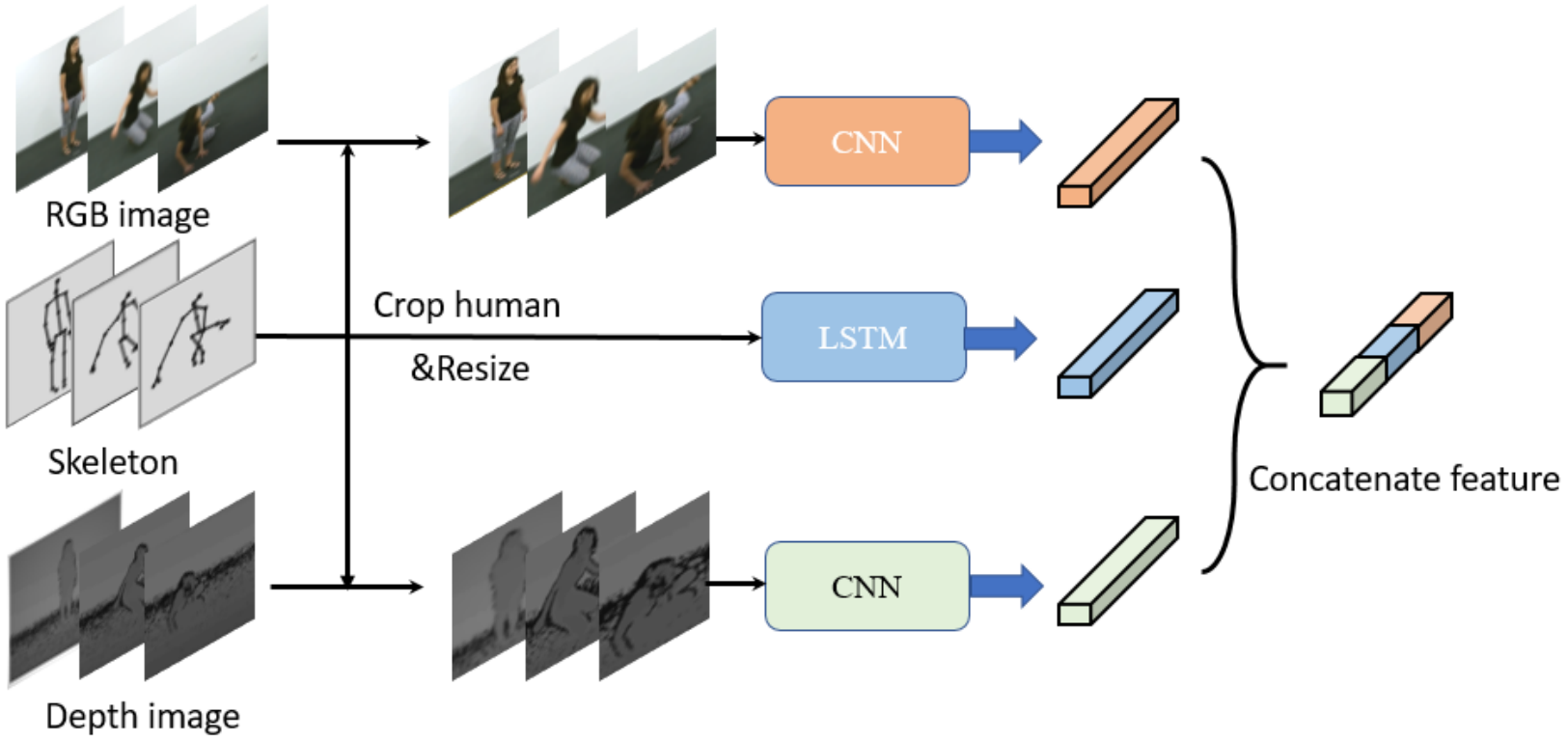
- Recognition: actions are fully executed and captured, **after-the-fact** analysis.
- Prediction: recognizing actions before they are fully executed, **pre-fact analysis**.

1. Introduction & Motivation



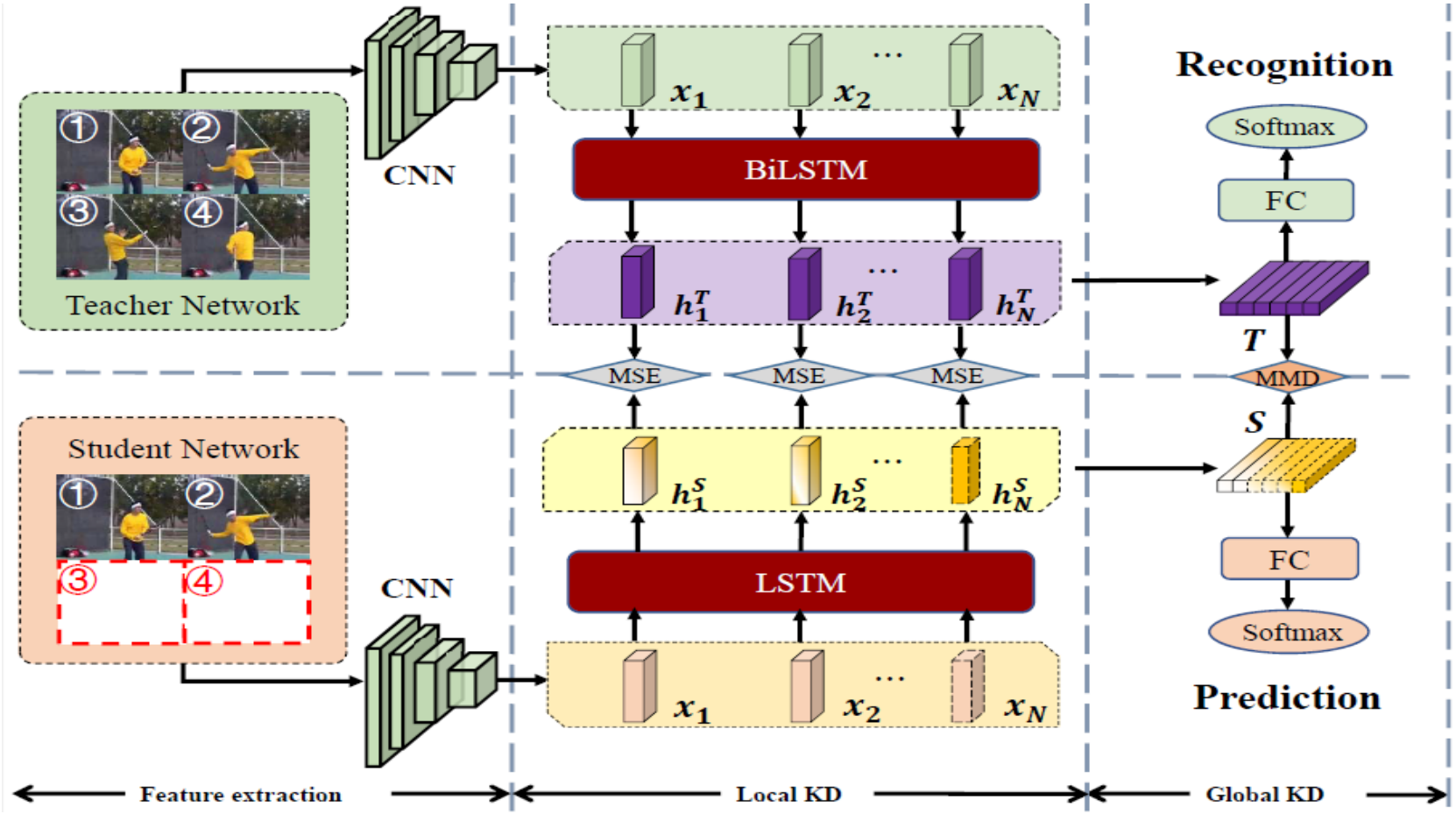
- Full videos contain **more action information** than partial videos.
- **Knowledge distilled from full videos** is beneficial for the action prediction with partially observed action videos.

2. Feature Extraction



- Extracting multimodal features (16-channel CNN and LSTM)
- Concatenating multimodal features along the modality dimension
- More features often means more useful action information

3. Teacher Student Learning



- Knowledge distillation cross two different video tasks
- Distilling both local and global knowledge for obtaining better prediction results

3. Teacher Student Learning

- **Local progressive-wise knowledge:**

$$L_{MSE}(S_i, T_i) = \|S_i \odot \omega - T_i \odot \omega\|_F^2$$

Mean Square Error loss: Each individual progress level

- **Global distribution knowledge:**

$$L_{MMD}(S_i, T_i) = \left\| \frac{1}{N} \sum^{n} \phi(S_i(:, n) - \frac{1}{N} \sum^{n} \phi(T_i(:, n) \right\|^2$$

Maximum Mean Discrepancy loss: ALL the progress levels

It can be simplified as:

$$L_{MMD}(S_i, T_i) = \|S_i S_i^{\top} - T_i T_i^{\top}\|_2^2$$

When ϕ is the second order polynomial kernel function:

$$k(x, y) =< \phi(x), \phi(y) >= (x^{\top} y)^2$$

3. Experiments

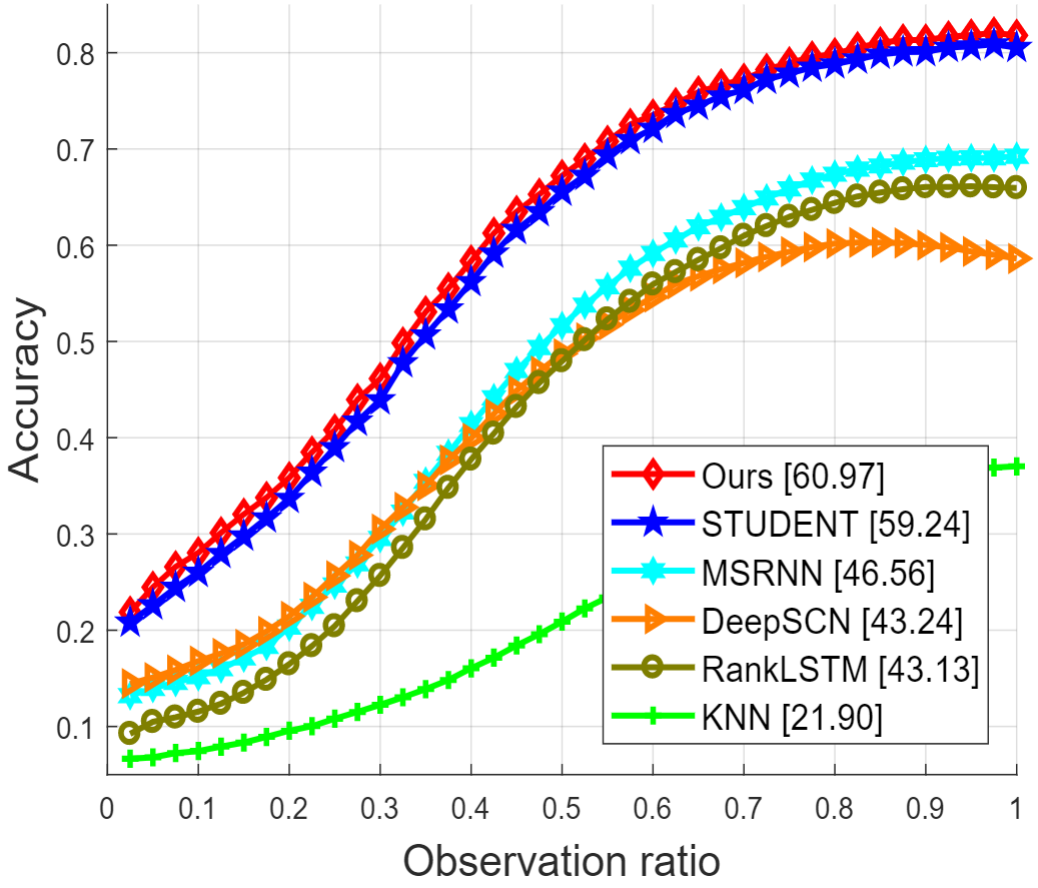
- **NTU RGB+D Dataset:**
  - 60 action classes, >56K action sequences, 3 camera views
  - 40 progress levels
  - RGB, depth, and skeleton modalities

Observation ratio	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	AUC
KNN [16]	7.45	9.56	12.25	16.04	20.89	25.97	30.85	34.49	36.15	37.02	21.90
RankLSTM [28]	11.54	16.48	25.66	37.74	47.96	55.94	60.99	64.41	66.05	65.95	43.13
DeepSCN [24]	16.80	21.46	30.51	39.93	48.73	54.61	58.18	60.18	60.01	58.62	43.24
MSRNN [16]	15.17	20.33	29.53	41.37	51.64	59.15	63.91	67.38	68.89	69.24	46.56
STUDENT	25.99	33.68	43.91	56.20	65.59	72.12	76.16	78.82	80.09	80.53	59.24
Ours	**27.80**	**35.85**	**46.27**	**58.45**	**67.40**	**73.86**	**77.63**	**80.06**	**81.47**	**82.01**	**60.97**

3. Experiments

- **NTU RGB+D Dataset:**
  - 60 action classes, >56K action sequences, 3 camera views
  - 40 progress levels
  - RGB, depth, and ske

Observation ratio	10%	20%	30%
KNN [16]	7.45	9.56	12.25
RankLSTM [28]	11.54	16.48	25.66
DeepSCN [24]	16.80	21.46	30.51
MSRNN [16]	15.17	20.33	29.53
STUDENT	25.99	33.68	43.91
Ours	**27.80**	**35.85**	**46.27**



3. Experiments

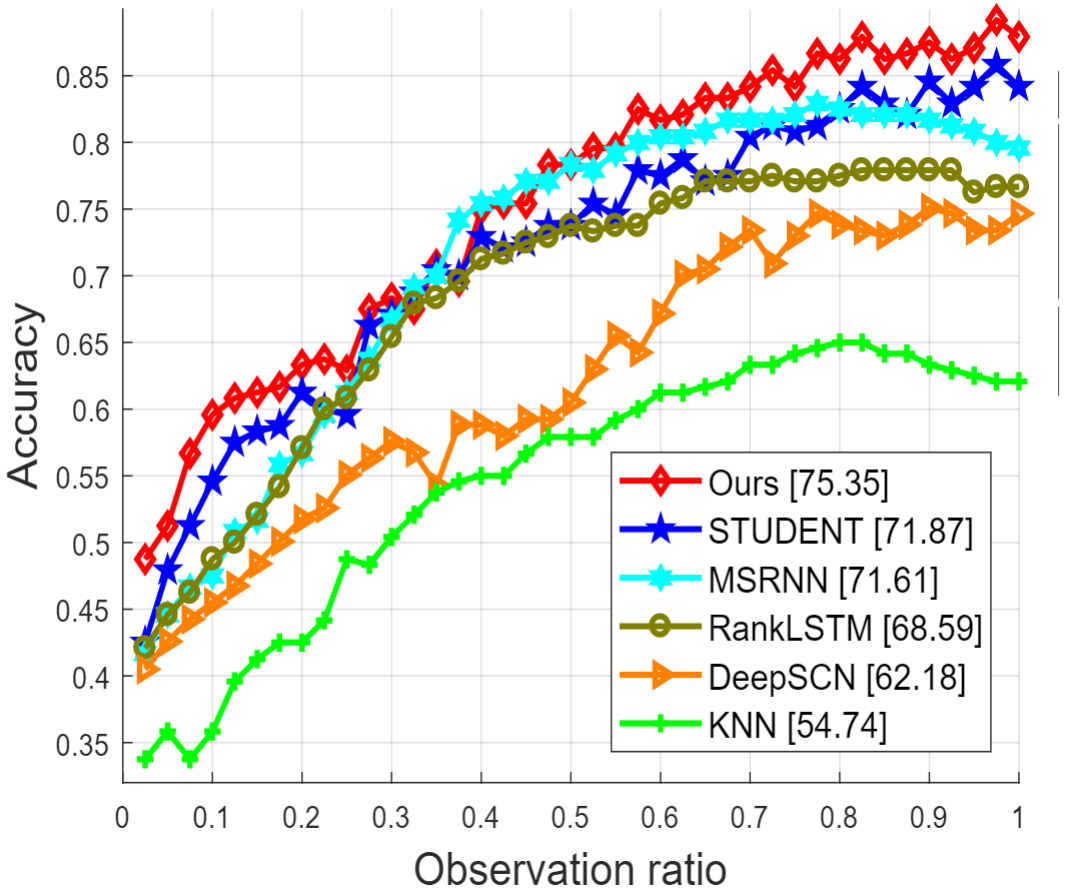
- **SYSU 3D HOI set:**
  - 12 action classes (6 pairs of HOI) , 480 action sequences
  - RGB, depth, and skeleton modalities
  - 40 progress level

Observation ratio	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	AUC
KNN [16]	35.83	42.50	50.42	55.00	57.92	61.25	63.33	65.00	63.33	62.08	54.74
RankLSTM [28]	48.75	57.08	65.42	71.25	73.75	75.42	77.08	77.50	77.92	76.67	68.59
DeepSCN [24]	45.50	51.75	57.58	58.83	60.50	67.17	73.42	73.83	75.08	74.67	62.18
MSRNN [16]	47.50	56.67	66.67	**75.42**	**78.33**	80.42	81.67	82.50	81.67	79.58	71.61
STUDENT	54.58	61.25	67.08	72.92	73.75	77.50	80.42	82.50	84.58	84.17	71.87
Ours	**59.58**	**63.33**	**68.33**	75.00	**78.33**	**81.67**	**84.17**	**86.25**	**87.50**	**87.92**	**75.35**

3. Experiments

- **SYSU 3D HOI set:**
  - 12 action classes (6 pairs of HOI) , 480 action sequences
  - RGB, depth, and skeleton modalities
  - 40 progress level

Observation ratio	10%	20%	30%	40
KNN [16]	35.83	42.50	50.42	55
RankLSTM [28]	48.75	57.08	65.42	71
DeepSCN [24]	45.50	51.75	57.58	58
MSRNN [16]	47.50	56.67	66.67	**75**
STUDENT	54.58	61.25	67.08	72
Ours	**59.58**	**63.33**	**68.33**	75



3. Experiments

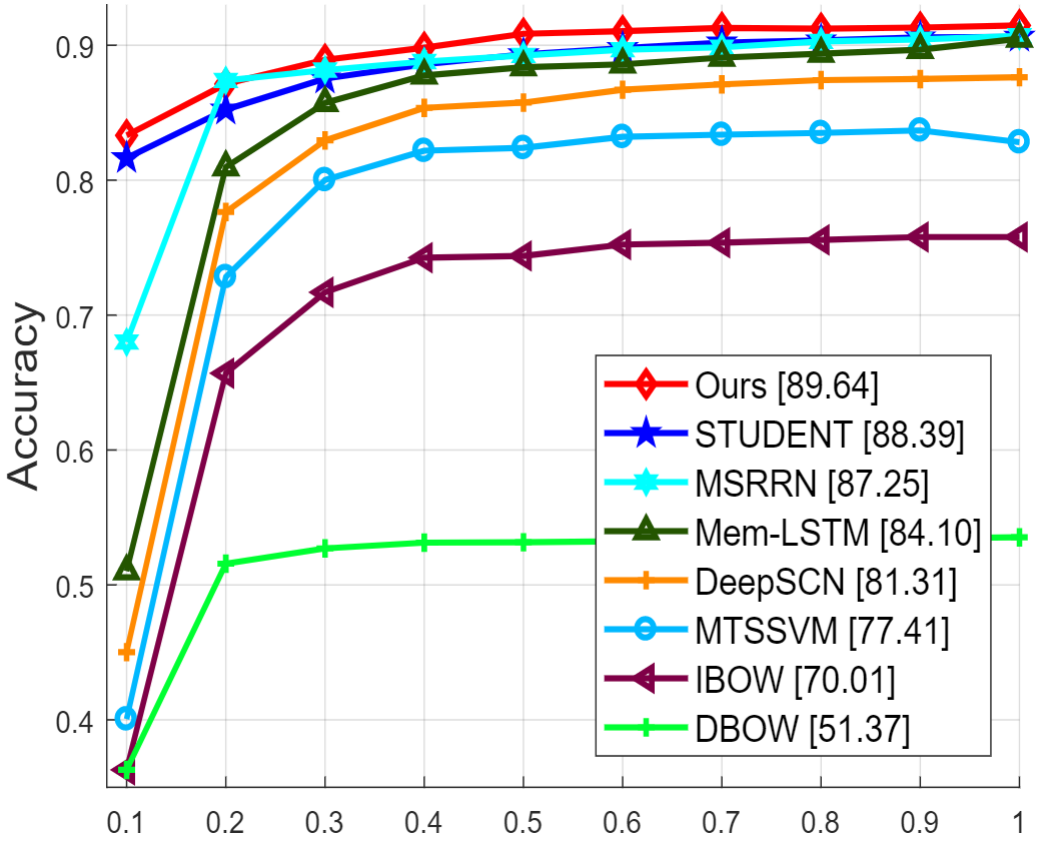
- **UCF-101 set:**
  - 101 action classes, >13K action sequences
  - Unconstrained video set (RGB)
  - 10 progress levels

Observation ratio	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	AUC
DBOW [32]	36.29	51.57	52.71	53.13	53.16	53.24	53.24	53.34	53.45	53.53	51.37
IBOW [32]	36.29	65.69	71.69	74.25	74.39	75.23	75.36	75.57	75.79	75.79	70.01
MTSSVM [23]	40.05	72.83	80.02	82.18	82.39	83.21	83.37	83.51	83.69	82.82	77.41
DeepSCN [24]	45.02	77.64	82.95	85.36	85.75	86.70	87.10	87.42	87.50	87.63	81.31
Mem-LSTM [22]	51.02	80.97	85.73	87.76	88.37	88.58	89.09	89.38	89.67	90.49	84.10
MSRNN [16]	68.00	**87.39**	88.16	88.79	89.24	89.67	89.85	90.28	90.43	90.70	87.25
STUDENT	81.64	85.23	87.53	88.59	89.33	89.79	90.20	90.36	90.58	90.63	88.39
Ours	**83.32**	87.13	**88.92**	**89.82**	**90.85**	**91.04**	**91.28**	**91.23**	**91.31**	**91.47**	**89.64**

3. Experiments

- **UCF-101 set:**
  - 101 action classes, >13K action sequences
  - Unconstrained video set (RGB)
  - 10 progress levels

Observation ratio	10%	20%	30%	40%
DBOW [32]	36.29	51.57	52.71	53.1
IBOW [32]	36.29	65.69	71.69	74.2
MTSSVM [23]	40.05	72.83	80.02	82.1
DeepSCN [24]	45.02	77.64	82.95	85.3
Mem-LSTM [22]	51.02	80.97	85.73	87.7
MSRNN [16]	68.00	**87.39**	88.16	88.7
STUDENT	81.64	85.23	87.53	88.5
Ours	**83.32**	87.13	**88.92**	**89.8**



3. Experiments

- **Ablation study:**
  - with vs. without MSE(L) and MMD(G) losses, S stands for STUDENT only.

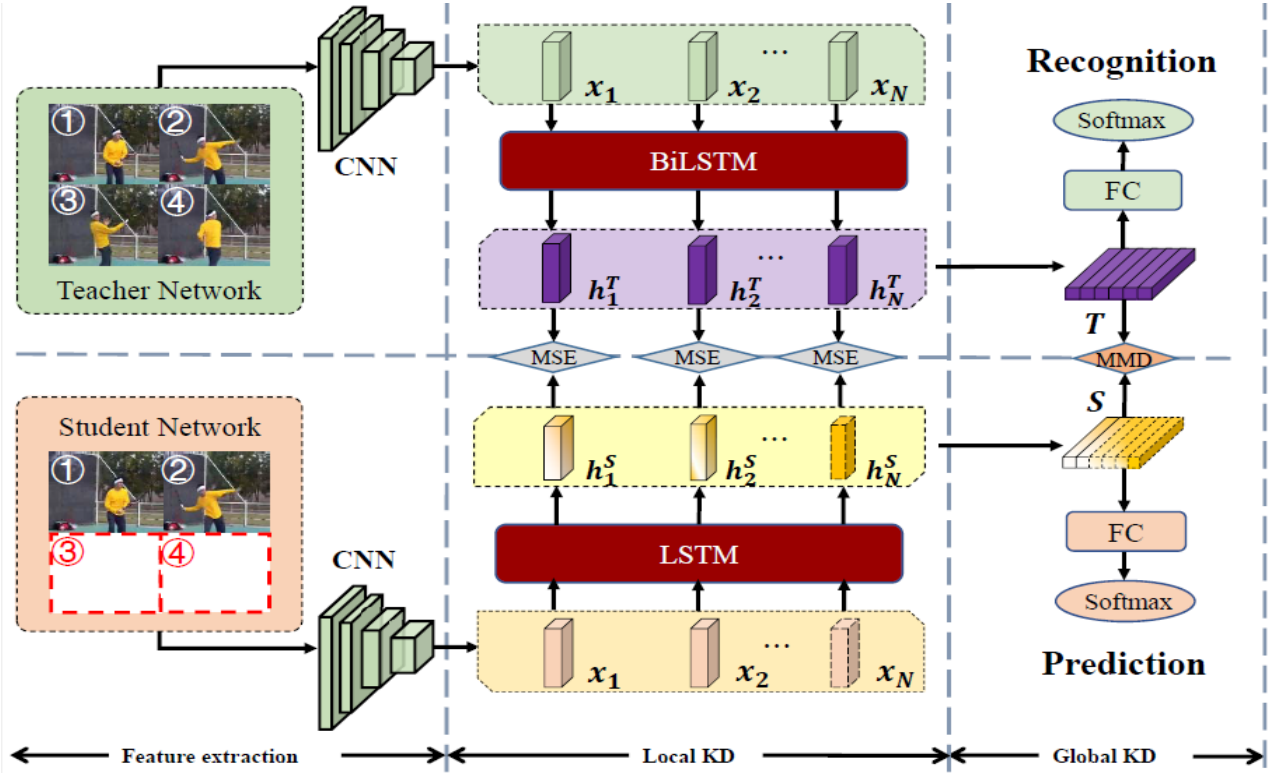
Observation ratio		10%	30%	50%	70%	100%	AUC
SYSU	S	54.58	67.08	73.75	80.42	84.17	71.87
	S+L	57.08	67.08	75.83	80.42	85.83	73.53
	S+G	57.50	66.67	76.67	80.42	85.00	73.08
	S+L+G	**59.58**	**68.33**	**78.33**	**84.17**	**87.92**	**75.35**
UCF-101	S	81.64	87.53	89.33	90.20	90.63	88.39
	S+L	83.19	88.43	90.22	91.20	90.98	89.27
	S+G	**83.57**	88.02	90.14	90.63	90.71	89.01
	S+L+G	83.32	**88.92**	**90.85**	**91.28**	**91.47**	**89.64**

  - With vs. without joint learning

Observation ratio		10%	30%	50%	70%	100%	AUC
SYSU	with	53.33	66.25	74.58	81.67	84.58	72.49
	without	**59.58**	**68.33**	**78.33**	**84.17**	**87.92**	**75.35**
UCF-101	with	**83.60**	88.35	89.82	90.20	90.85	89.07
	without	83.32	**88.92**	**90.85**	**91.28**	**91.47**	**89.64**

4. Summary

- A novel teacher-student learning framework for distilling progressive action knowledge from action recognition model(teacher) to early action prediction model(student);
- An early action prediction system integrating the action prediction task with action recognition in the spirit of knowledge distillation.



5. MORE INFO.

http://isee.sysu.edu.cn/~hujianfang

EMAIL ME: hujianf@mail2.sysu.edu.cn

感谢各位老师和同学！