Progressive Relation Learning for Group Activity Recognition

Guyue Hu Bo Cui Yuan He Shan Yu
Institute of Automation, Chinese Academy of Sciences

CVPR SEATTLE WASHINGTON JUNE 16-18 2020

1. Introduction

□Task: Group activity recognition

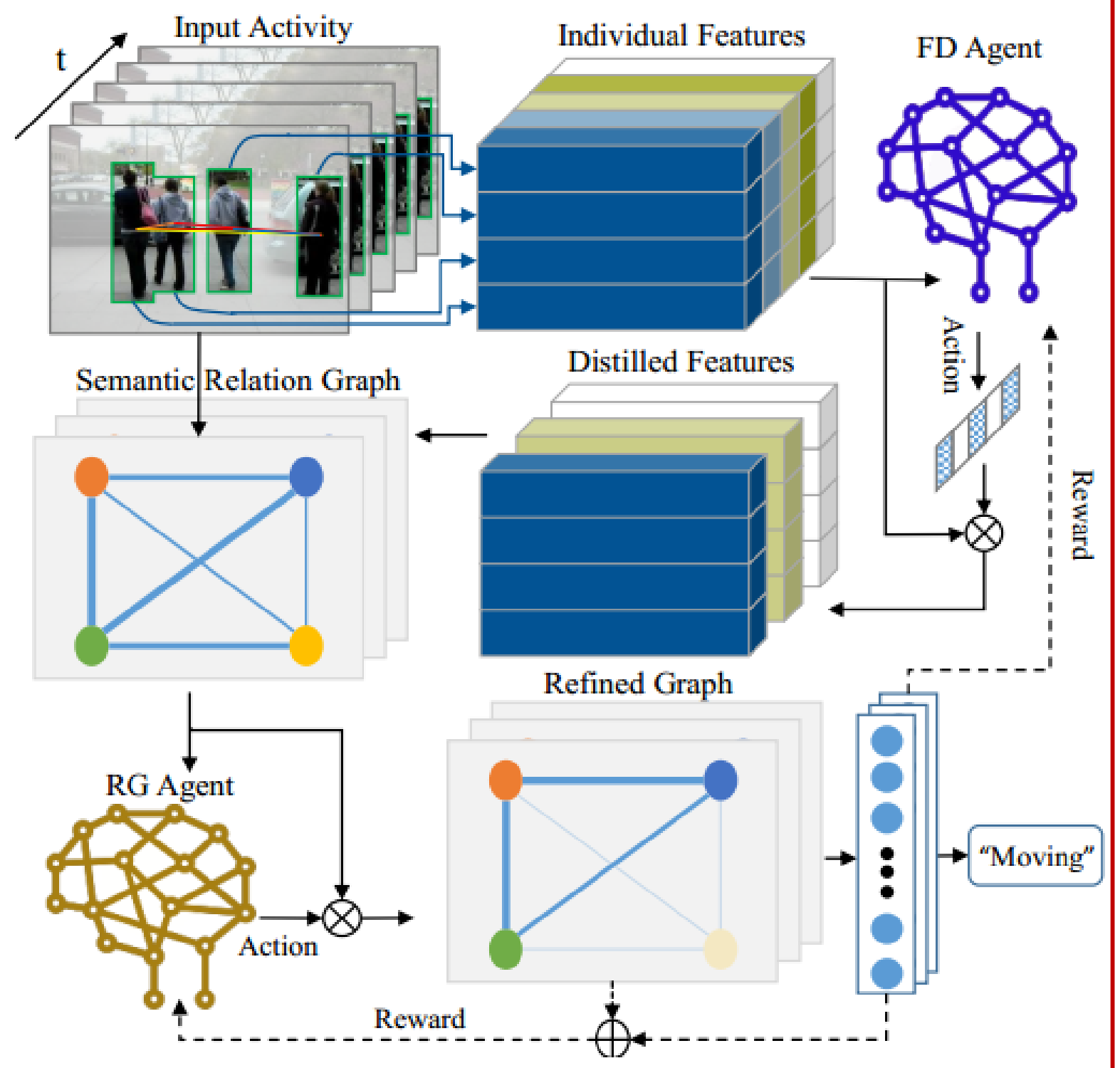
- **Input:** Videos containing many interactive individuals (persons).
- **Output:** Activity label of group behavior.

□Background

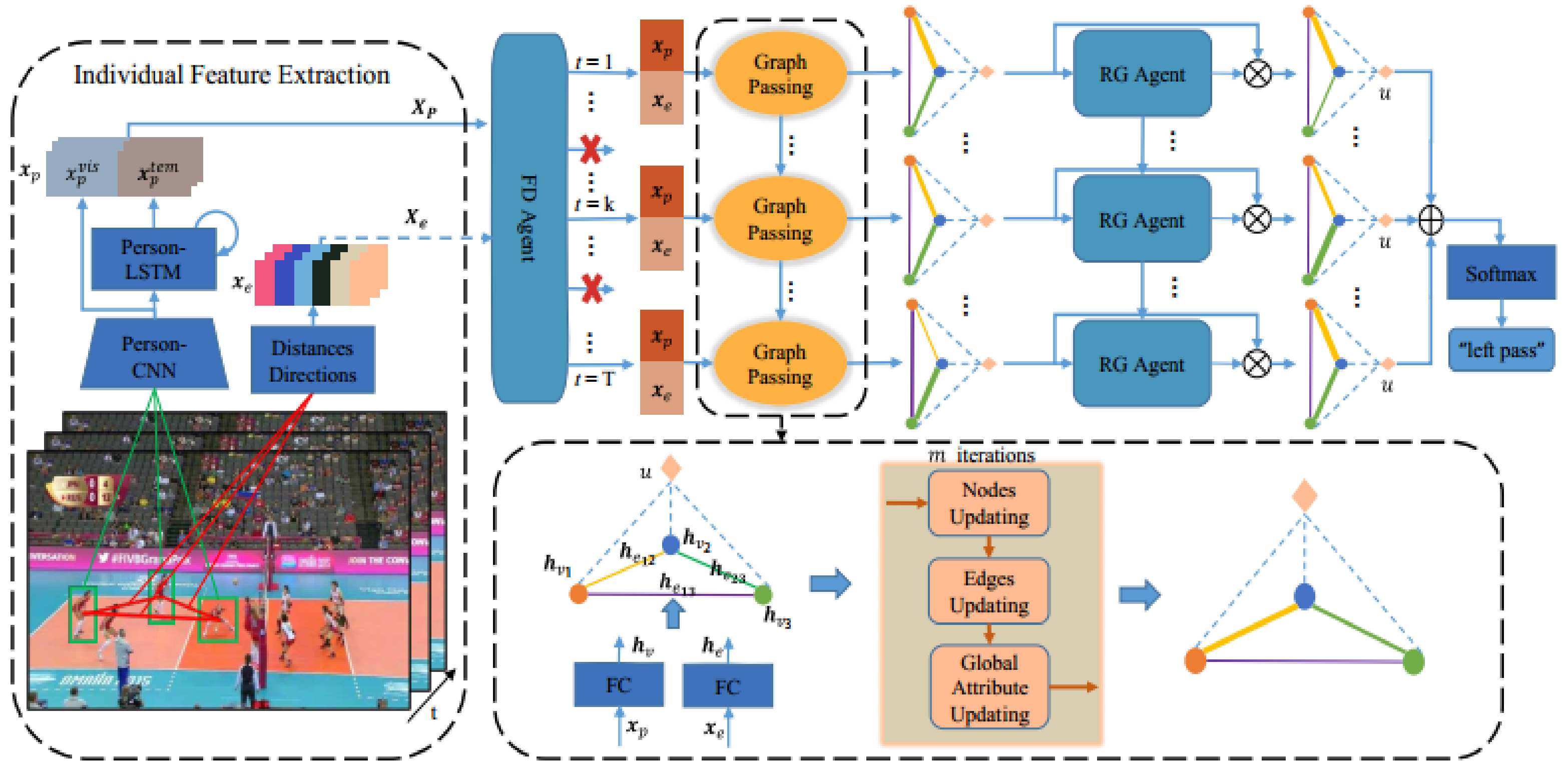
- Group activities involve dynamics among many interactive and noisy individuals.
- Only a few participants at several key frames dominate and finally define the group activity.

□Motivation

- **A semantic relation graph (SRG):** Model relations among individuals.
- **A Feature distilling (FD) agent:** Refine low-level individual features by distilling informative frames.
- **A Relation-gating (RG) agent:** Adjust high-level semantic graph to attend to group-relevant relations.



2. Framework



□Individual Feature Extraction

- Tracklests: tracked from person annotations in the middle frames.
- Extract the individual spatiotemporal features (*Xp*) and the original interaction features (*Xe*).

□Semantic Relation Graph

- Node (person) attributes are initialized as individual spatiotemporal feature, edge attributes are initialized as original interaction feature, *u* represents global attribute (e.g., activity score)
- The graph updated for *m* iterations during each forward pass.

□ RL-based agents

- Two agents adopting policy according to two Markov decision processes are proposed to progressively refine the graph.
- The structure and decision process are in the next two sections.

3. RG Agent

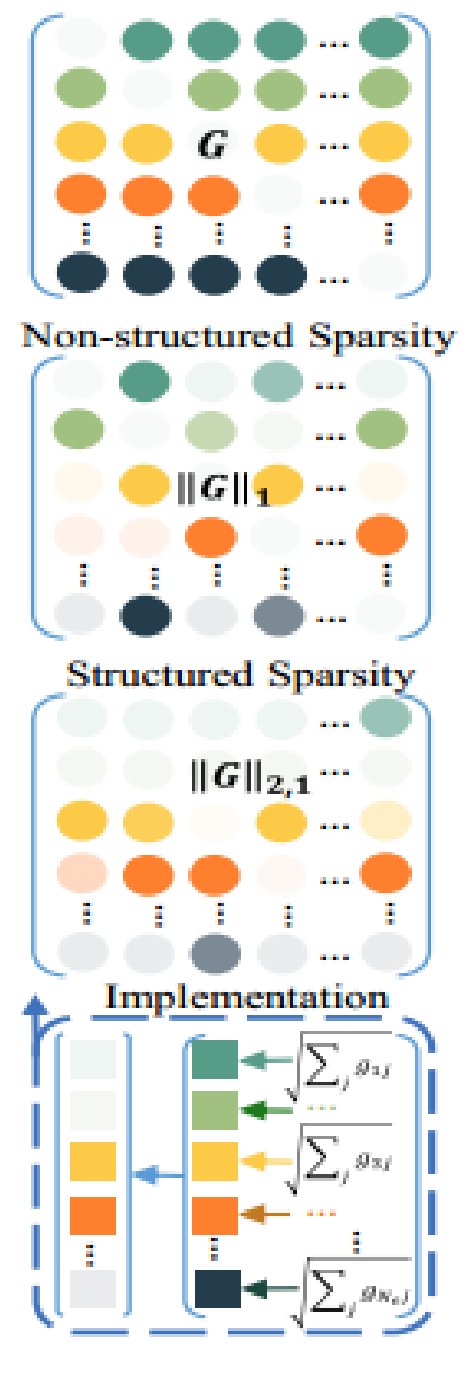
□Action

- Generate a gate for each relation

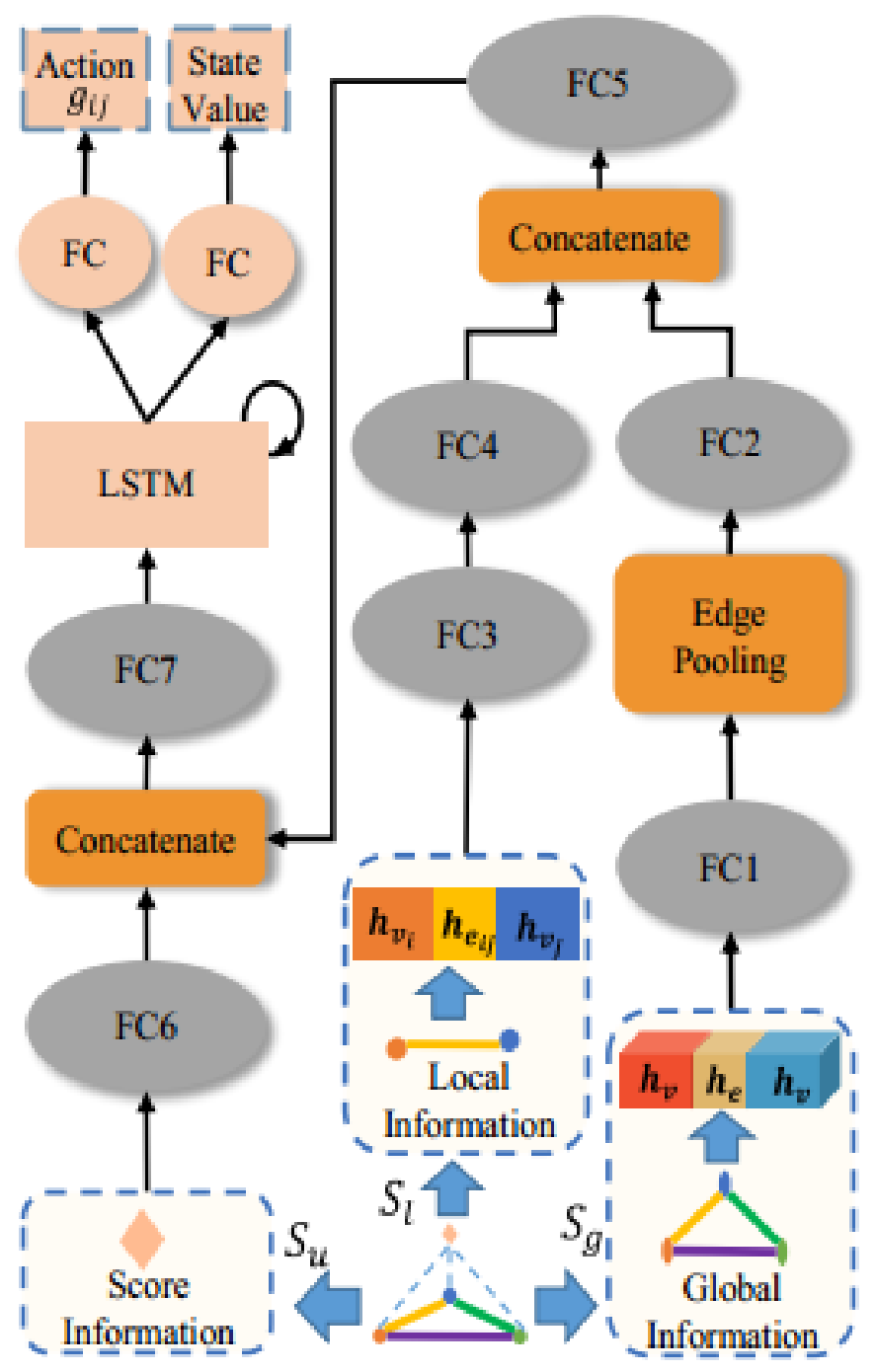
$$\boldsymbol{h}_{e_{ij}} = \boldsymbol{h}_{e_{ij}} \cdot g_{ij}$$

□State

- S_g: the whole semantic graph, i.e., the stack of all relation triplets (sender, relation， receiver").
- S_l: local relation triplet.
- S_u: activity score.



(a) Comparison of sparsity



(b) Structure of the RG agent

□Reward:

- **Sparse reward:** attend to a few key participants and relations,

$$r_{sparse} = -sgn(L_{2,1}(\boldsymbol{G}_\tau) - L_{2,1}(\boldsymbol{G}_{\tau-1}))$$

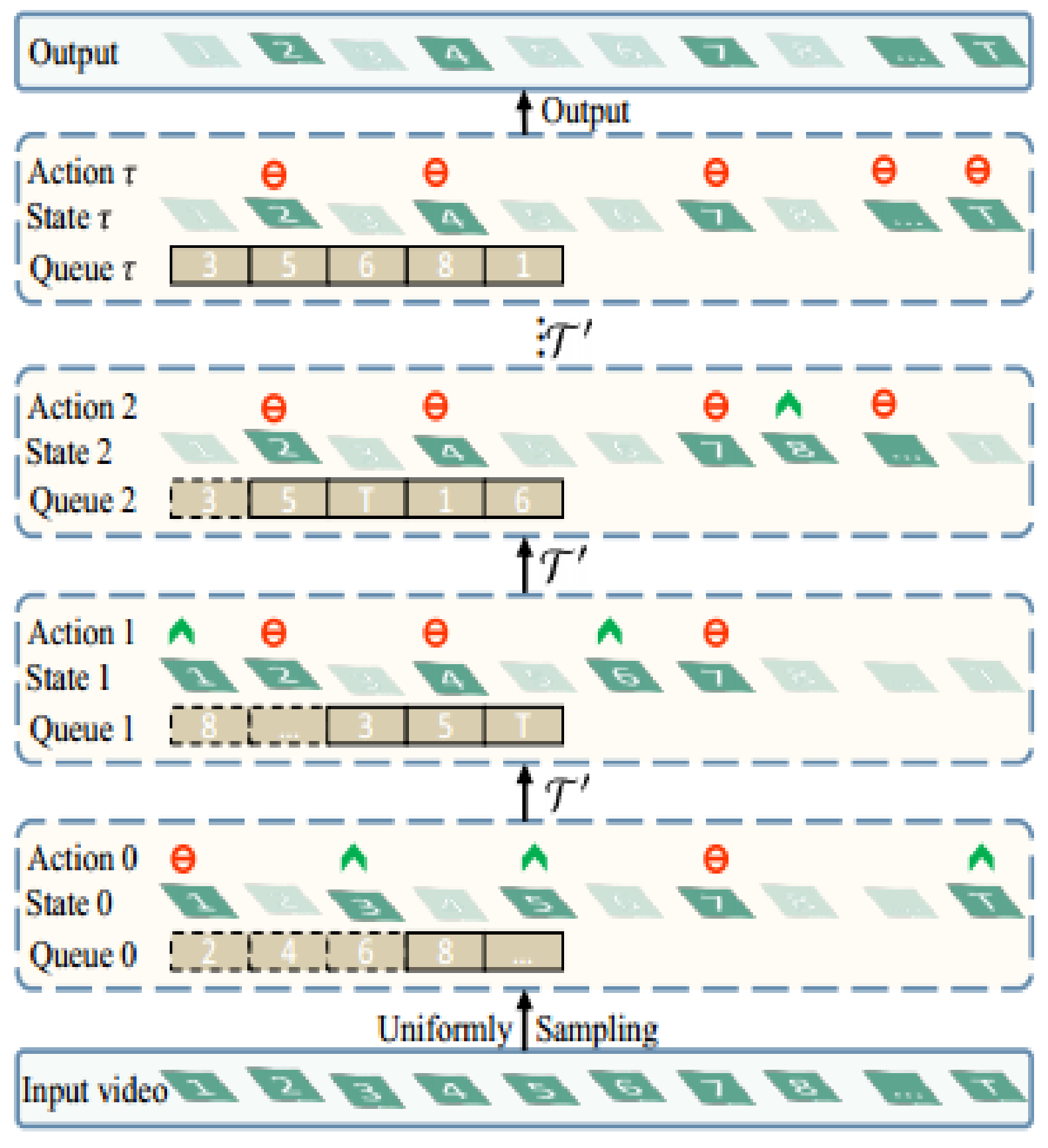
- **Ascend reward**: encourage to evolve along an ascending trajectory,

$$r_{ascend} = sgn(\boldsymbol{p}_\tau^c - \boldsymbol{p}_{\tau-1}^c)$$

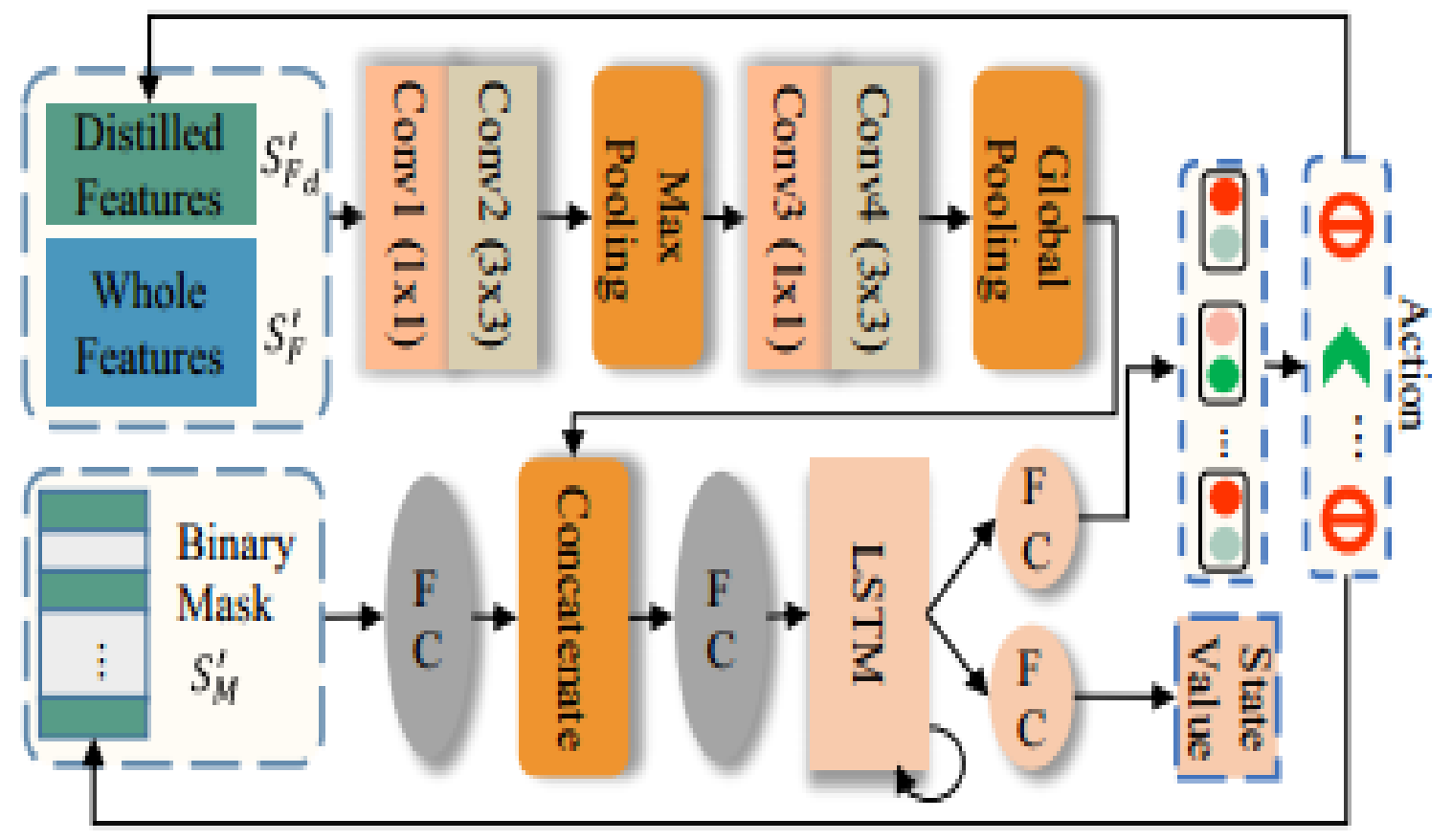
- **Shift reward**: enforce a strong stimulation/punishment Ω/-Ω when class shifting after a reinforcement step.
- **In total**, the reward for RG agent is

$$r = r_{sparse} + r_{ascend} + r_{shift}$$

4. FD Agent



(a) Illustration of the feature-distilling process



(b) Structure of the feature-distilling agent

□State

- S'_F: whole feature tensor of an group activity.
- S'_{F_d}: feature of the distilled frames (implicit local info).
- S'_M: binary mask of distilled frames (explicit local info).

□Action

Generate two types of discrete action for each selected frame:

- **"stay distilled"** indicating the frame is informative that the agent determines to keep it.
- **"shift to alternate"** indicating the agent determines to discard the frame and take in an alternate.

□Reward

- Contain the two components about trajectory ascending and class shifting introduced above, i.e.,

$$r = r_{ascend} + r_{shift}.$$

5. Training

□Alternate Training

- Totally 9 separated training stages.
- At each stage, only one of the three components (SRG, FD Agent, RG Agent) is trained and the remaining two are frozen (or removed).
- Individual features are extracted and saved to disk previously, thus only need reloading in these stages.
- Two agents are both optimized with the classical A3C algorism.

6. Experiment

□Accuracy Comparison

Volleyball

Methods	Backbone	OF	MCA	MPCA
HDTM [14]	AlexNet	N	81.9	82.9
SBGAR [19]	Inception-v3	Y	66.9	67.6
CERN-2 [25]	VGG16	N	83.3	83.6
SSU [2]	Inception-v3	N	89.9	-
SRNN [4]	AlexNet	N	83.5	-
PC-TDM [36]	AlexNet	Y	87.7	88.1
stagNet [22]	VGG16	N	89.3	-
SPA+KD [31]	VGG16	N	89.3	89.0
SPA+KD+OF [31]	VGG16	Y	90.7	90.0
ARG [33]	VGG16	N	**91.9**	-
CRM [1]	I3D	Y	**93.0**	-
Baseline [22]	VGG16	N	87.9	-
Ours-SRG	VGG16	N	88.3	88.5
Ours-SRG+T. A.	VGG16	N	88.6	88.7
Ours-SRG+R. A.	VGG16	N	88.7	89.0
Ours-SRG+FD	VGG16	N	89.5	89.2
Ours-SRG+RG	VGG16	N	89.8	91.1
Ours-PRL	VGG16	N	**91.4**	**91.8**

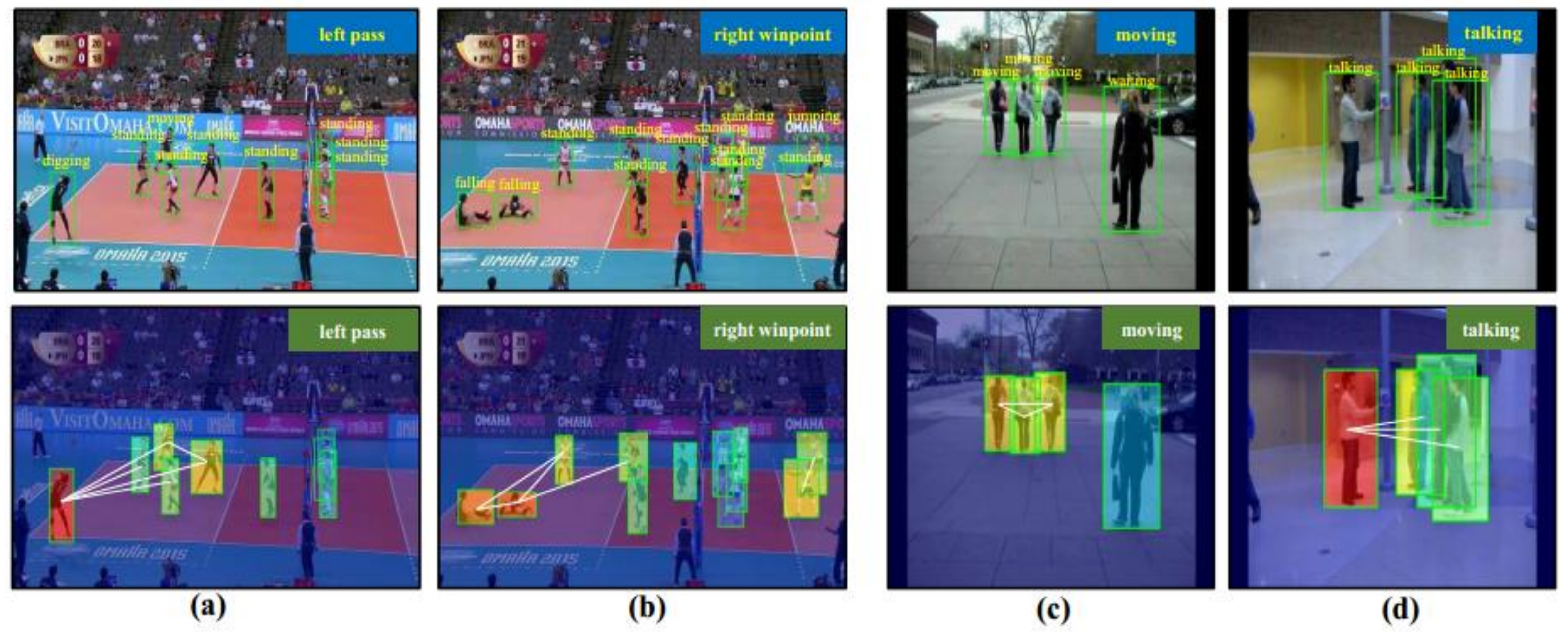
CAD

Methods	Backbone	OF	MPCA(%)
HDTM [14]	AlexNet	N	89.6
CERN-2 [25]	VGG16	N	88.3
SBGAR [19]	Inception-v3	Y	89.9
PC-TDM [36]	AlexNet	Y	92.2
SPA+KD [31]	VGG16	N	92.5
SPA+KD+OF [31]	VGG16	Y	**95.7**
CRM [1]	I3D	Y	94.2
Baseline [22]	VGG16	N	87.7*
Ours-SRG	VGG16	N	89.4
Ours-SRG+R. A.	VGG16	N	90.0
Ours-SRG+T. A.	VGG16	N	90.1
Ours-SRG+FD	VGG16	N	91.1
Ours-SRG+RG	VGG16	N	91.4
Ours-PRL	VGG16	N	**93.8**

* MPCA is unavailable, MCA is listed instead.

- The three components SRG, RG Agent, and FD- Agent are effective.
- Progressive relation learning is superior to attention variants.

□Visualization Result



Visualization of the refined SRGs. Color: importance degree of person. White lines: relations with top5/top3 (Volleyball/CAD) gate values.

- Discover the subset of relations related to the "digging" person is the key to determine the activity "left pass".
- Predict "right winpoint" mainly based on two relation clusters, i.e., the "falling cluster"(left) and "cheering cluster"(right).
- Concentrate on the relations among the three moving persons to suppress the noisy relations caused by the "Waiting" person.
- Attend to the relations connected to the "Talking" person and weakens the relations among the three audiences.