Temporal Action Localization with Weak Supervision and Language

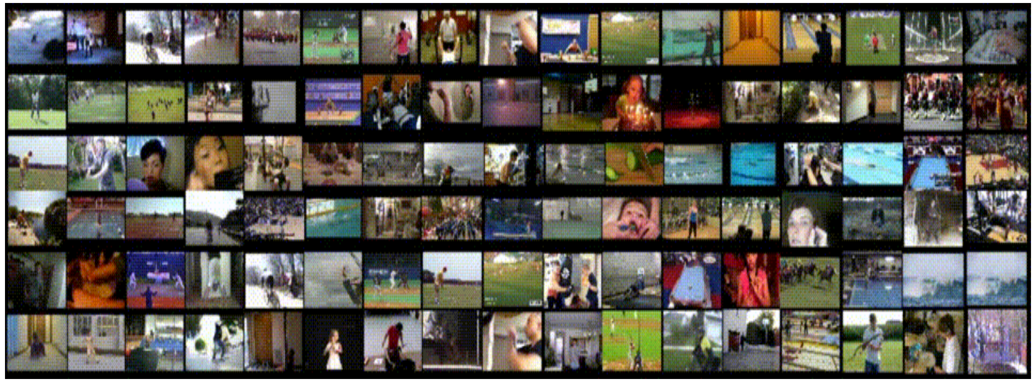
Yan Huang

Center for Research on Intelligent Perception and Computing (CRIPAC)
National Laboratory of Pattern Recognition (NLPR)
Institute of Automation, Chinese Academy of Sciences (CASIA)

Nov. 8, 2020

Background

- **Action Classification** (video classification)
  - trimmed video
  - predict an action label



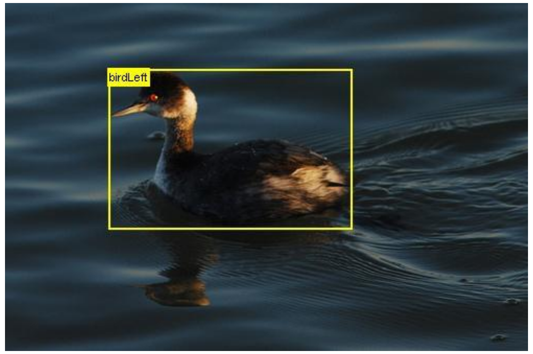
- **Action Localization** (temporal action localization)
  - untrimmed video
  - predict intervals and labels of actions

Longboarding

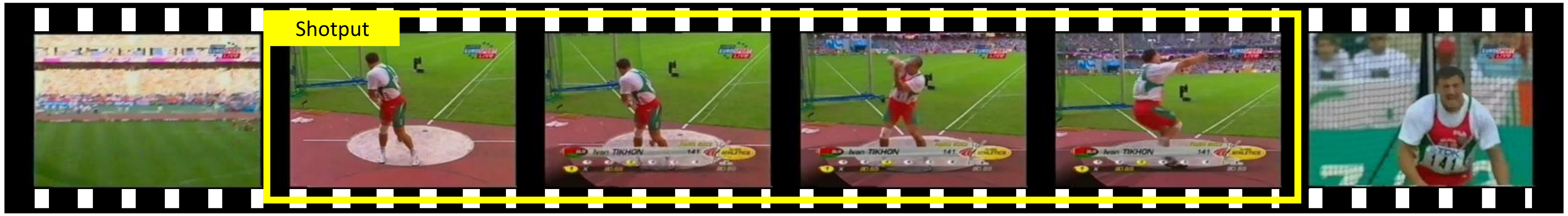
Tumbling

Annotations

Expensive



Object proposal
(one glimpse)



Temporal action proposal
(multiple glimpses)

Subjective on action boundaries

The research community has been interested in weakly-supervised temporal action localization (WTAL)

Other Weak Supervisions

Only class annotation

No temporal boundary

Longboarding

Language annotation

Multiple actions

A person runs to the window and then look out

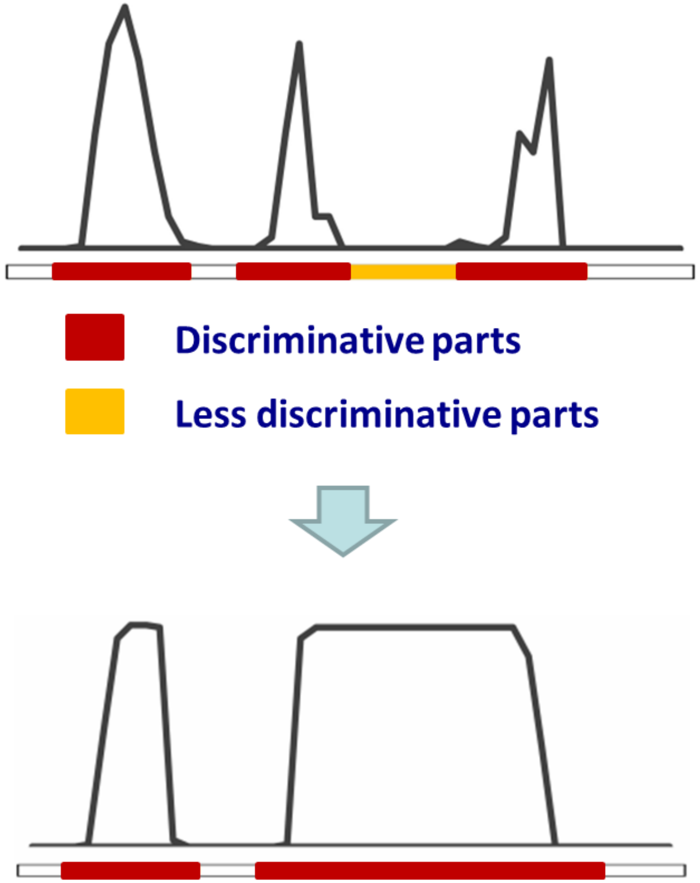
Relational Prototypical Network for Weakly Supervised Temporal Action Localization

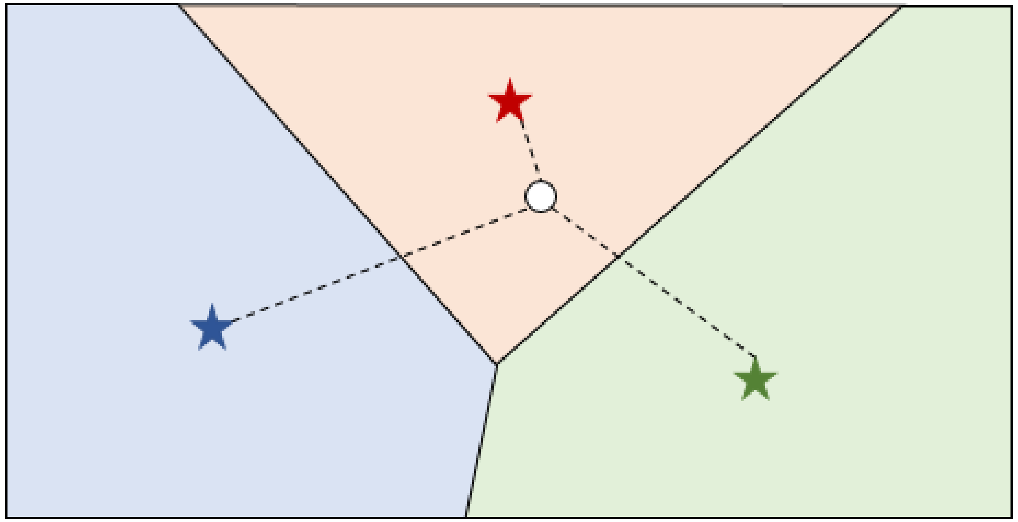
Linjiang Huang, Yan Huang , Wanli Ouyang, Liang Wang

AAAI 2020 (Oral)

Our Motivation

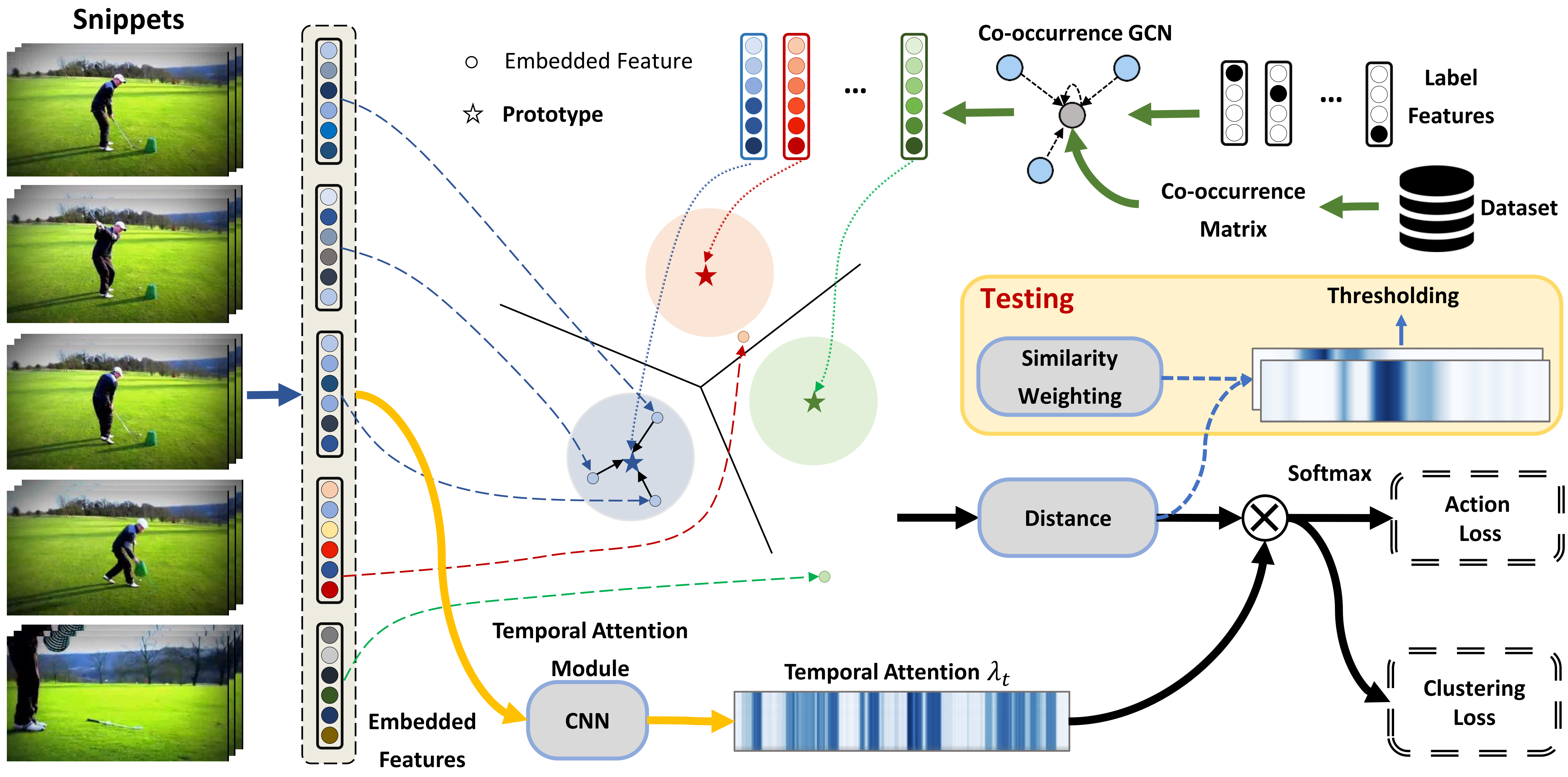
- **Learning compact and discriminative features is difficult, due to the imbalance distribution of different actions**
- **Modeling relations among actions with prototypical network**



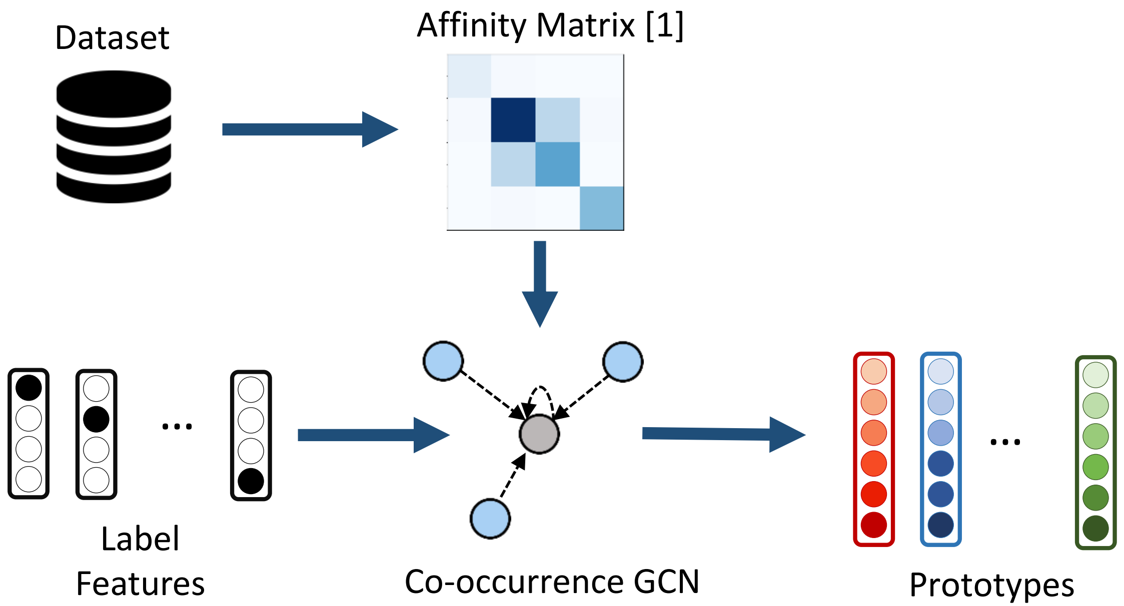


Prototypical network

Model Architecture



Prototype Embedding Module



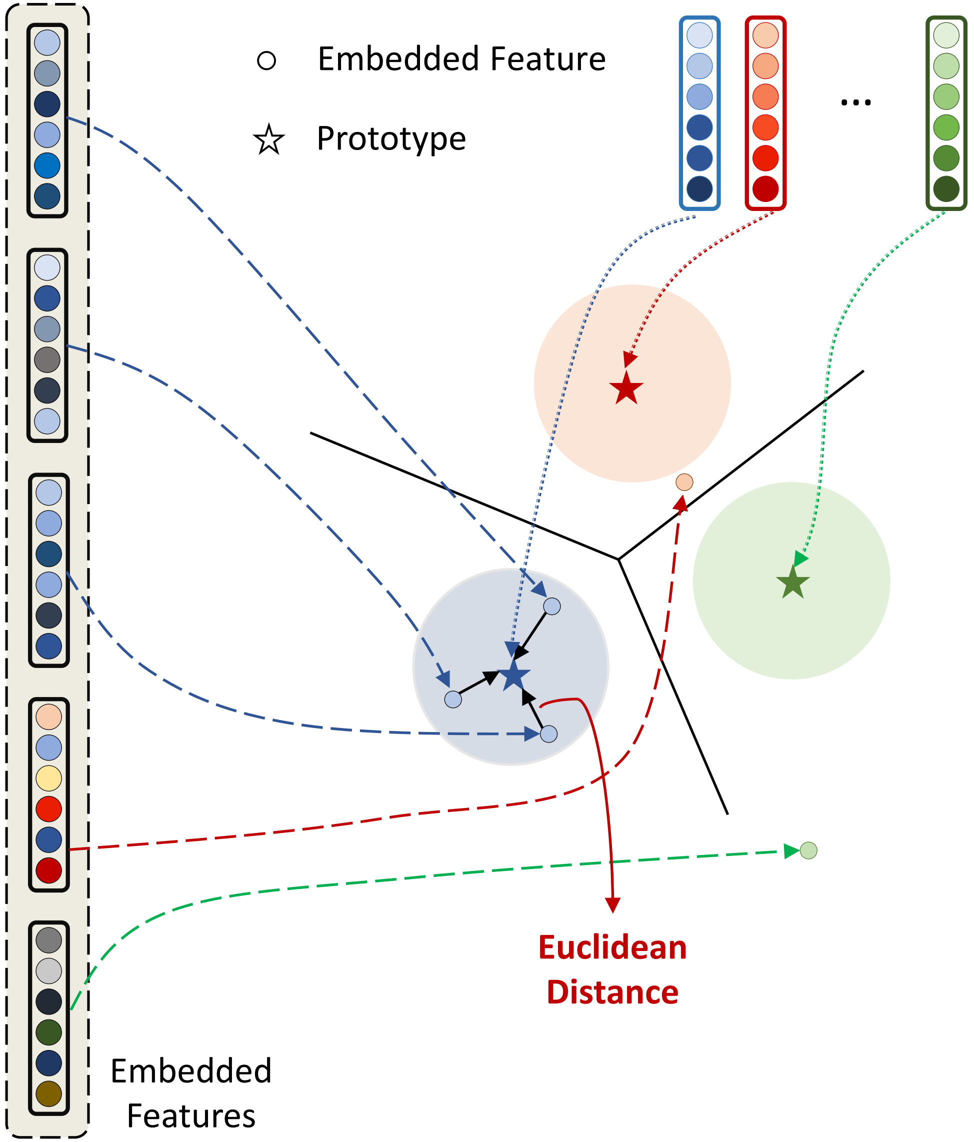
- Affinity matrix derived from statistics in the dataset
- Label features represent different actions
- Co-occurrence GCN captures relations and pulls related prototypes closer

Learning inter-dependent prototypes rather than independent prototypes

Prototypes $\boldsymbol{P} = \{\boldsymbol{p}_i\}_{i=1}^{C}$

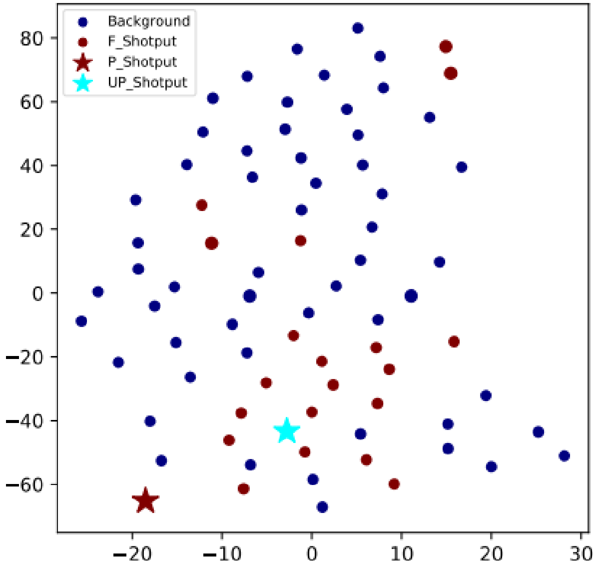
[1] Chen, Z.-M.; Wei, X.-S.; Wang, P.; and Guo, Y. Multi-label image recognition with graph convolutional networks, CVPR 2019
[2] Li, Q.; Han, Z.; and Wu, X.-M. 2018. Deeper insights into graph convolutional networks for semi-supervised learning. AAAI 2018

Prototype Matching Module

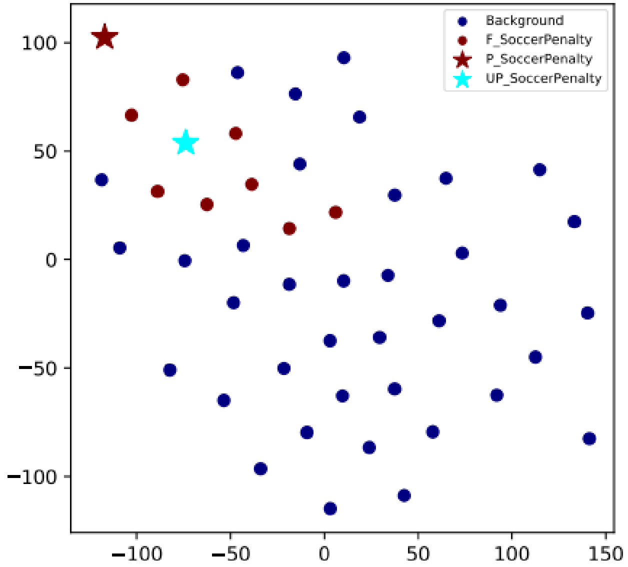


Negative Euclidean distance is employed as the similarity between feature and prototype

Matching score $s_{tj} = -\|\boldsymbol{x}_e^t - \boldsymbol{p}_j\|_2^2$



t-SNE on the video of shotput



t-SNE on the video of soccerpenalty

Comparison with SOTA

Detection performance comparisons over the THUMOS14 dataset.

Supervision	Method	AP @ IoU							
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	AVG (0.1:0.5)
Full	S-CNN (Shou et al. 2016)	47.7	43.5	36.4	28.7	19.0	-	5.3	35.0
Full	R-C3D (Xu et al. 2017)	54.5	51.5	44.8	35.6	28.9	-	-	43.1
Full	SSN (Zhao et al. 2017)	**60.3**	56.2	50.6	40.8	29.1	-	-	47.4
Full	TAL-Net (Chao et al. 2018)	59.8	**57.1**	**53.2**	**48.5**	**42.8**	**33.8**	**20.8**	**52.3**
Weak	Hide-and-Seek (Singh et al. 2018)	36.4	27.8	19.5	12.7	6.8	-	-	20.6
Weak	UntrimmedNet (Wang et al. 2017a)	44.4	37.7	28.2	21.1	13.7	-	-	29.0
Weak	SbS Erasion (Zhong et al. 2018)	45.8	39.0	31.1	22.5	15.9	-	-	30.9
Weak	STPN (UNT) (Nguyen et al. 2018)	45.3	38.8	31.1	23.5	16.2	9.8	5.1	31.0
Weak	W-TALC (UNT) (Paul et al. 2018)	49.0	42.8	32.0	26.0	18.8	-	6.2	33.7
Weak	AutoLoc (UNT) (Shou et al. 2018)	-	-	35.8	29.0	**21.2**	13.4	5.8	-
Weak	CMCS (UNT) (Liu et al. 2019)	53.5	46.8	37.5	29.1	19.9	12.3	6.0	37.4
Weak	Ours (UNT)	**54.2**	**47.1**	**37.8**	**29.4**	**21.2**	**13.9**	**6.8**	**37.9**
Weak	STPN (I3D) (Nguyen et al. 2018)	52.0	44.7	35.5	25.8	16.9	9.9	4.3	35.0
Weak	W-TALC (I3D) (Paul et al. 2018)	55.2	49.6	40.1	31.1	22.8	-	7.6	39.8
Weak	CMCS (I3D) (Liu et al. 2019)	57.4	50.8	41.2	32.1	23.1	15.0	7.0	40.9
Weak	MAAN (I3D) (Yuan et al. 2019)	59.8	50.8	41.1	30.6	20.3	12.0	6.9	40.5
Weak	Ours (I3D)	**62.3**	**57.0**	**48.2**	**37.2**	**27.9**	**16.7**	**8.1**	**46.5**

0.5% ↑

6.0% ↑

Comparison with SOTA

Table 2: Results on ActivityNet1.2 validation set. The AVG indicates the average mAP at IoU thresholds 0.5:0.05:0.95.

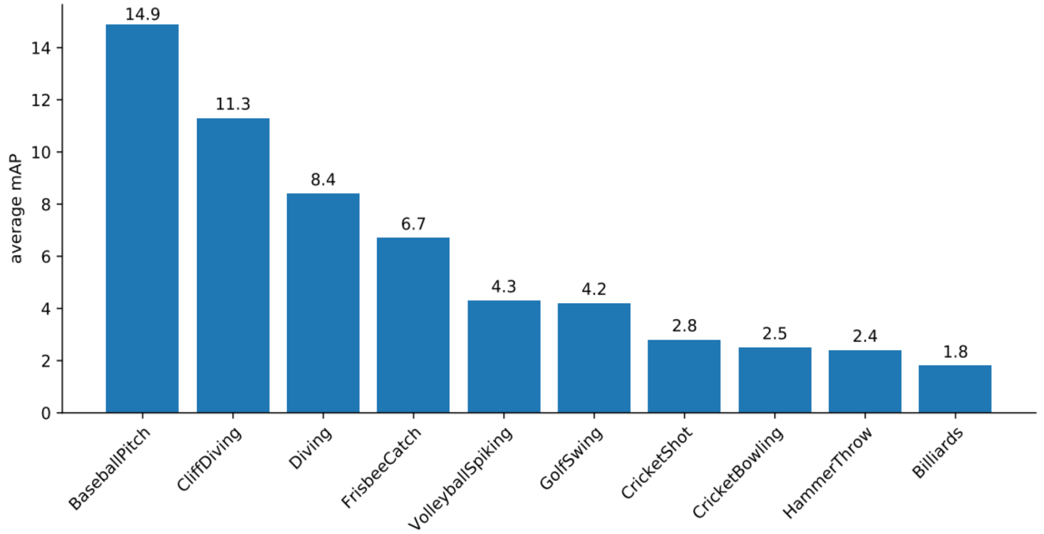
Method	AP @ IoU			
	0.5	0.75	0.95	AVG
Step-by-Step Erasion	27.3	14.7	2.9	15.6
AutoLoc (U)	27.3	15.1	3.3	16.0
CMCS (U)	33.9	19.9	5.1	20.5
Ours (U)	**37.0**	**21.1**	**5.2**	**22.0**
W-TALC (I)	37.0	-	-	18.0
CMCS (I)	36.8	22.0	**5.6**	22.4
Ours (I)	**37.6**	**23.9**	5.4	**23.3**

1.5% ↑

0.9% ↑

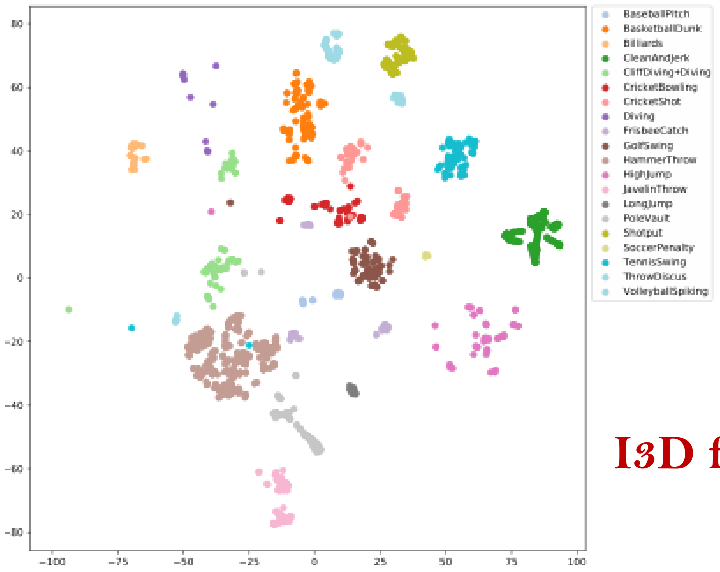
Table 3: Ablation study on prototype embedding module.

Methods	Random Initialization	FC	GCN
AVG (0.1:0.5)	44.6	44.7	**46.5**



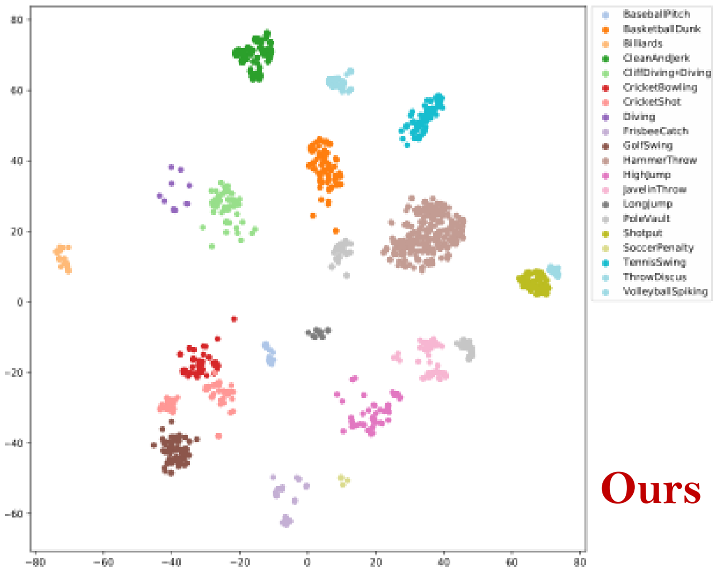
Class-specific gain resulting from building relations of actions. Performance differences between our full model and the one with random initialization are shown.

Evaluate the Learned Features



(a) t-SNE on I3D features

I3D features



(b) t-SNE on learned features

Ours

Method	AP @ IoU				
	0.1	0.3	0.5	AVG	Δ
SimpleNet	52.5	36.7	19.0	36.5	-
SimpleNet (our feature)	58.5	44.7	25.9	43.7	7.2
STPN (reported)	52.0	35.5	16.9	35.0	-
STPN (reproduced)	53.6	39.0	23.2	39.1	-
STPN (our feature)	58.0	42.9	25.2	42.4	3.3
W-TALC (reported)	55.2	40.1	22.8	39.8	-
W-TALC (reproduced)	55.2	40.3	23.7	40.0	-
W-TALC (our feature)	55.0	39.4	24.0	39.7	-0.3

- Our method indeed learns more compact features compare to the original I3D features
- The learned features can substantially improve the performance of temporal action localization

Language-Driven Temporal Activity Localization: A Semantic Matching Reinforcement Learning Model

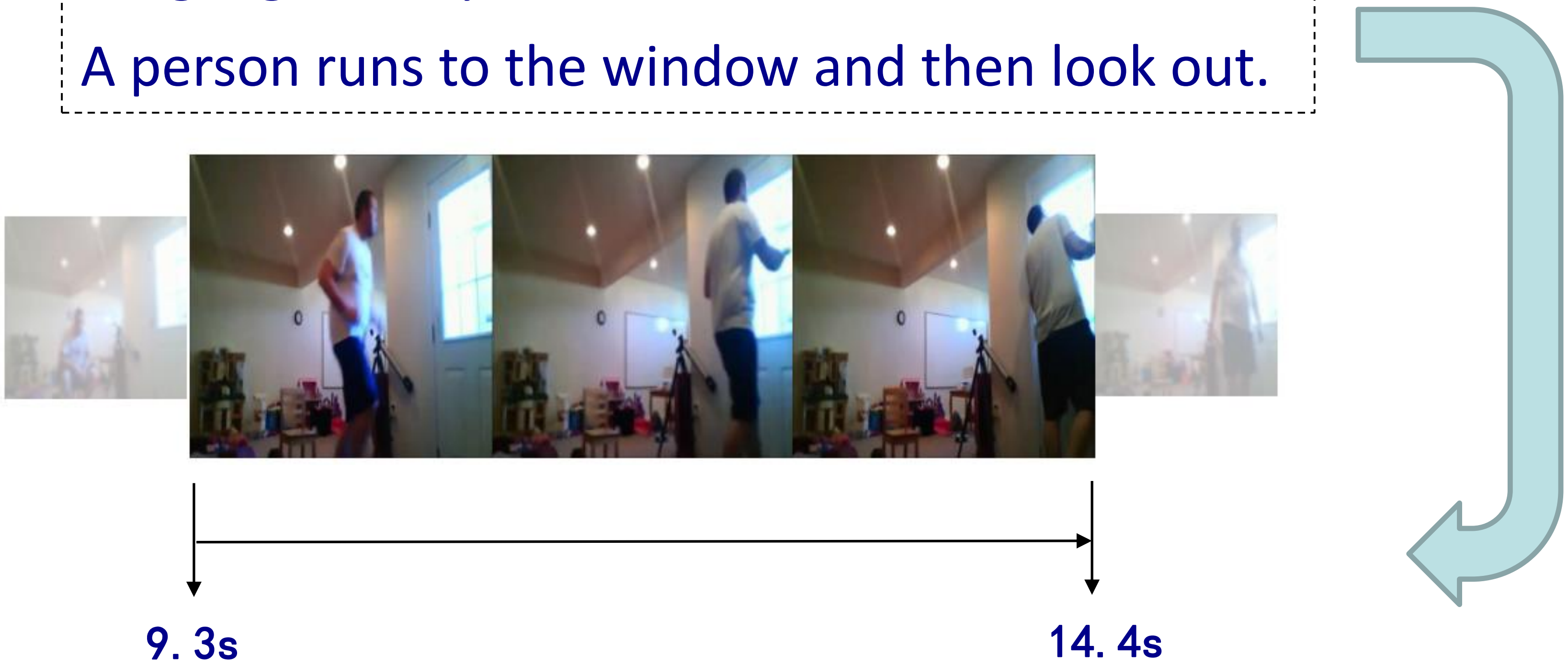
Weining Wang, Yan Huang , Liang Wang

CVPR 2019 (Oral)

Language-driven Temporal Action Localization

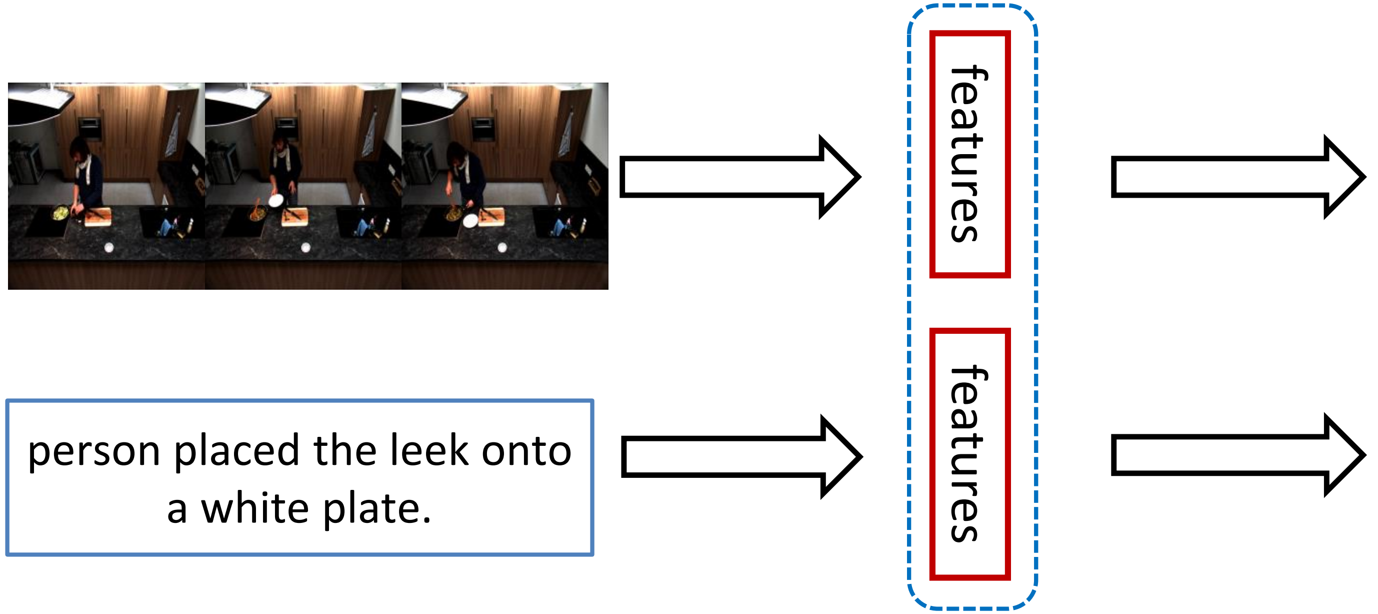
Language Query:

A person runs to the window and then look out.



Activities in real world are more complex and diverse, which cannot be well described by a single word

Our Motivation



- Alignment + Location Regression
  - TALL [1]
  - Attentive Moment Retrieval [2]
- Ranking Loss
  - Moment Context Network [3]

➢ Current methods are time-consuming with sliding windows

➢ Temporal information is not fully exploited

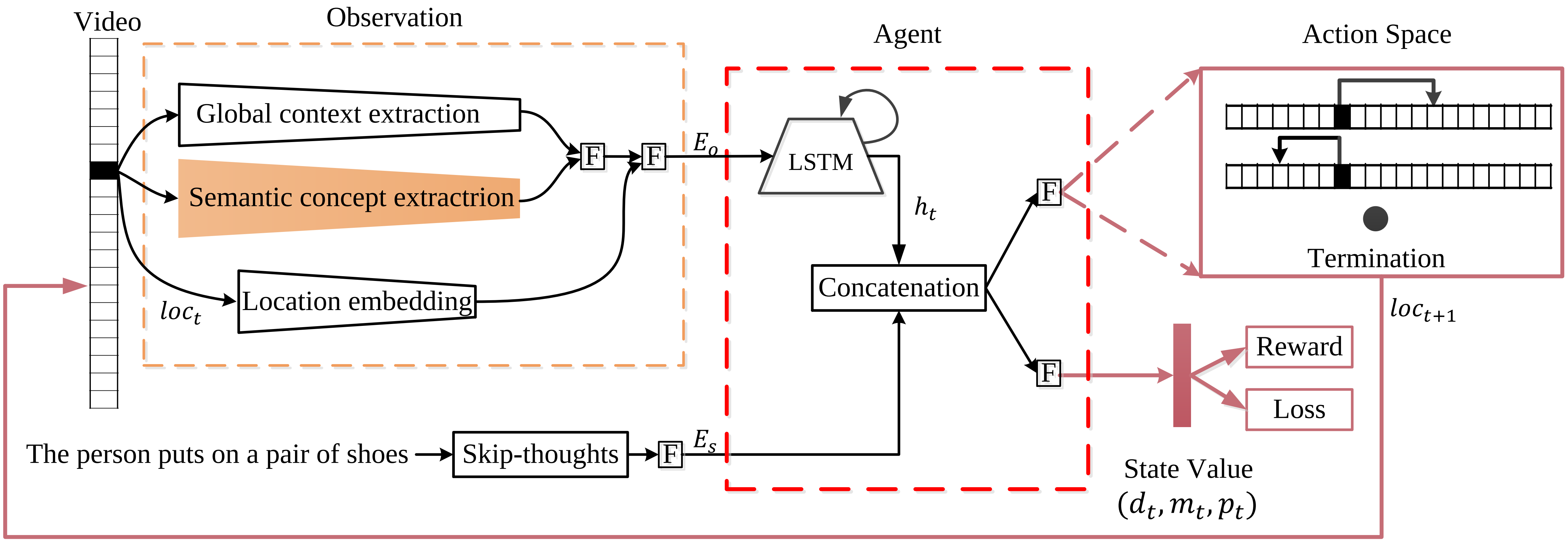
[1] Gao et al., TALL: Temporal Activity Localization via Language Query. In ICCV, 2017.

[2] Liu et al., Attentive moment retrieval in videos. In ACM SIGIR, 2018.

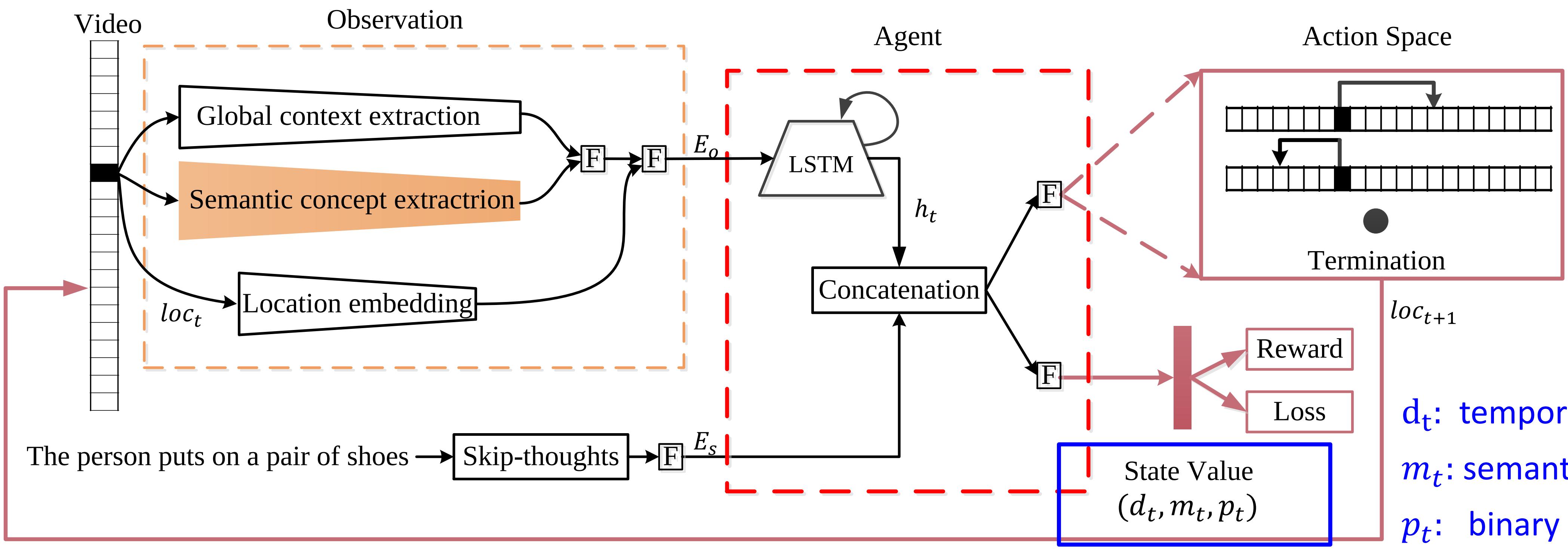
[3] Hendricks et al., Localizing moments in video with natural language. In ICCV, 2017..

Semantic Matching Reinforcement Learning

- **Regulate the temporal boundaries by selectively observing a sequence of video frames**
- **Match the visual-semantic information with the aid of semantic concepts**



Matching via Reinforcement Learning



d_t: temporal boundary prediction

m_t: semantic matching score

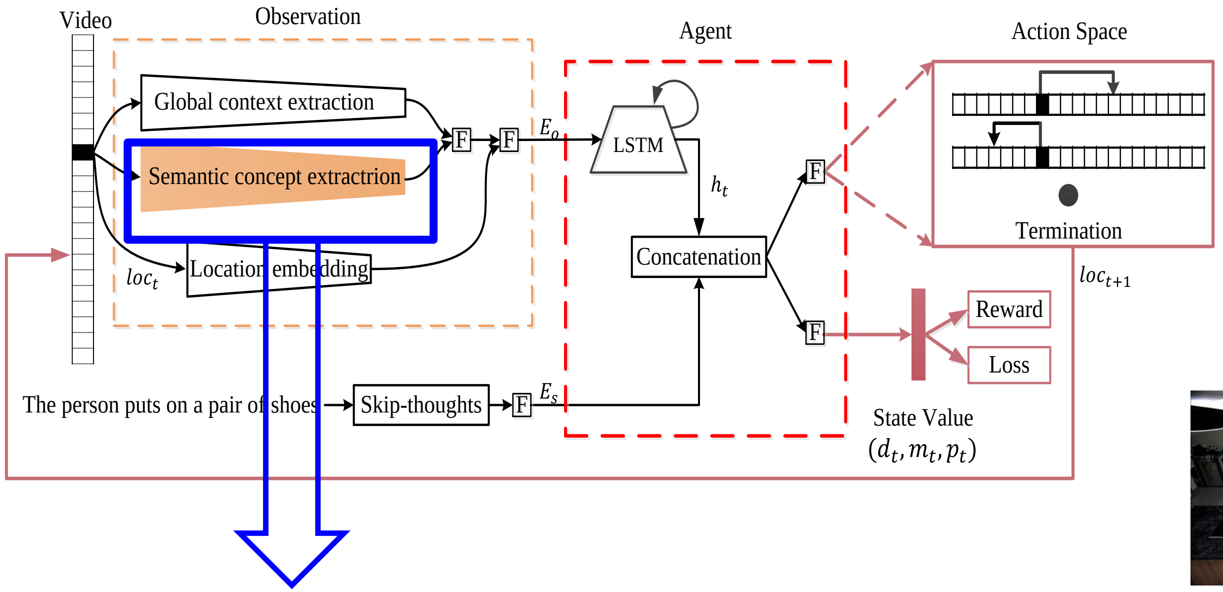
p_t: binary prediction indicator

$$L(D) = \omega_1 \sum_t L_{cls}(m_t) + \omega_2 \sum_t L_{loc}(d_t, g_t)$$

$$\begin{cases} L_{cls}(m_t; \theta_m) = -\sum_i (r_i) \log P(r_i | m_i; \theta_m) \\ L_{loc} = \frac{1}{n} \sum_i^n \sum_j^M [x_{ij} \log(p_{ij}) + (1 - x_{ij}) \log(1 - p_{ij})] \end{cases}$$

$$r_T = \begin{cases} R_{FN}, & FN \\ N_{TP} R_{TP} + N_{FP} R_{FP}, & TP \text{ and } FP \end{cases}$$

Sematic Concept Extraction



Semantic concepts of video frame

concept	score
person	1
orange	0.75
cut	0.93
...	
cutting board	0.83

The person cuts oranges on a cutting board

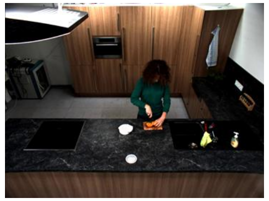
⇩

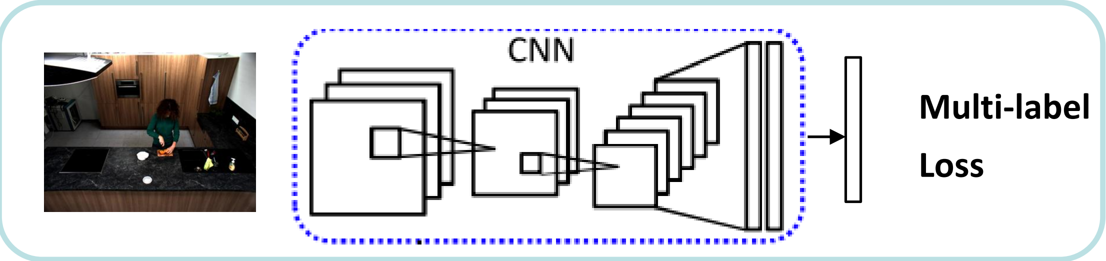
The, person, cuts, oranges, on, a, cutting board

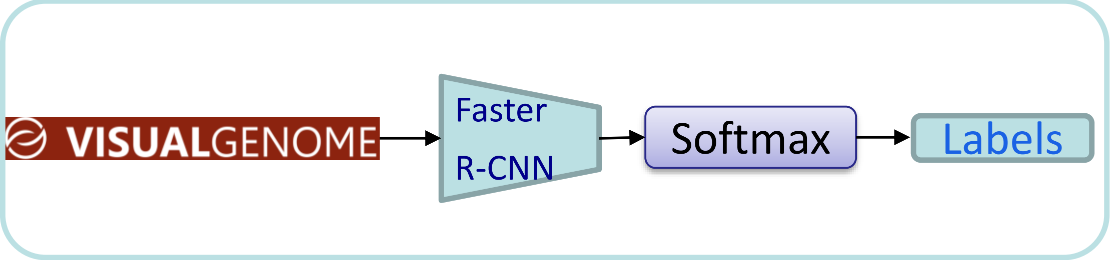
⇩

~~The~~, person, cuts, oranges, ~~on~~, ~~a~~, cutting board

⇩

 ⇔ person, cuts, oranges, cutting board





Experimental Results

Table 1: Results on TACoS and Charades-STA datasets

Method	TACoS							Charades-STA				
	R@1 IoU=0.5	R@1 IoU=0.3	R@1 IoU=0.1	R@5 IoU=0.5	R@5 IoU=0.3	R@5 IoU=0.1	mR	R@1 IoU=0.5	R@1 IoU=0.7	R@5 IoU=0.5	R@5 IoU=0.7	mR
Random	0.83	1.81	3.28	3.57	7.03	15.09	5.27	8.51	3.03	37.12	14.06	15.68
CTRL [8]	13.30	18.32	24.32	25.42	36.69	48.73	27.80	23.63	8.89	58.92	29.52	30.24
RL(b)	11.76	17.70	22.42	22.61	33.24	45.10	25.47	19.78	5.60	55.65	25.07	26.53
RL(f)	12.79	18.53	23.87	24.56	35.30	47.64	27.15	21.18	7.33	56.01	27.85	28.09
SM-RL(attr+b)	13.50	18.83	23.72	24.01	34.19	46.56	26.80	21.00	7.63	57.25	28.06	28.49
SM-RL(attr+f)	14.01	19.02	23.96	24.55	36.42	47.14	27.51	22.54	8.56	58.95	29.74	29.95
SM-RL(attr*+b)	14.20	19.79	25.17	25.38	36.69	48.22	28.24	23.56	9.52	60.17	32.53	31.45
SM-RL(attr*+f)	**15.95**	**20.25**	**26.51**	**27.84**	**38.47**	**50.01**	**29.84**	**24.36**	**11.17**	**61.25**	**32.08**	**32.22**

Table 2: Results on DiDeMo dataset

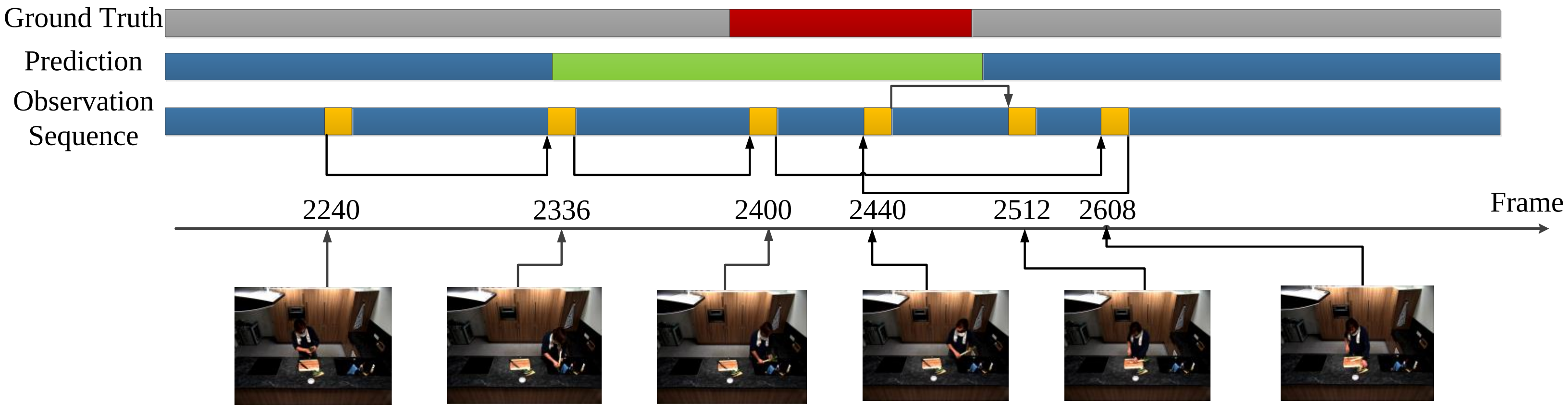
Method	Rank@1	Rank@5	mIoU
MCN([11])	28.10	78.21	41.08
SM-RL(attr*+b)	29.64	79.38	42.17
SM-RL(attr*+f)	**31.06**	**80.45**	**43.94**

Table 3: Detection speed comparison

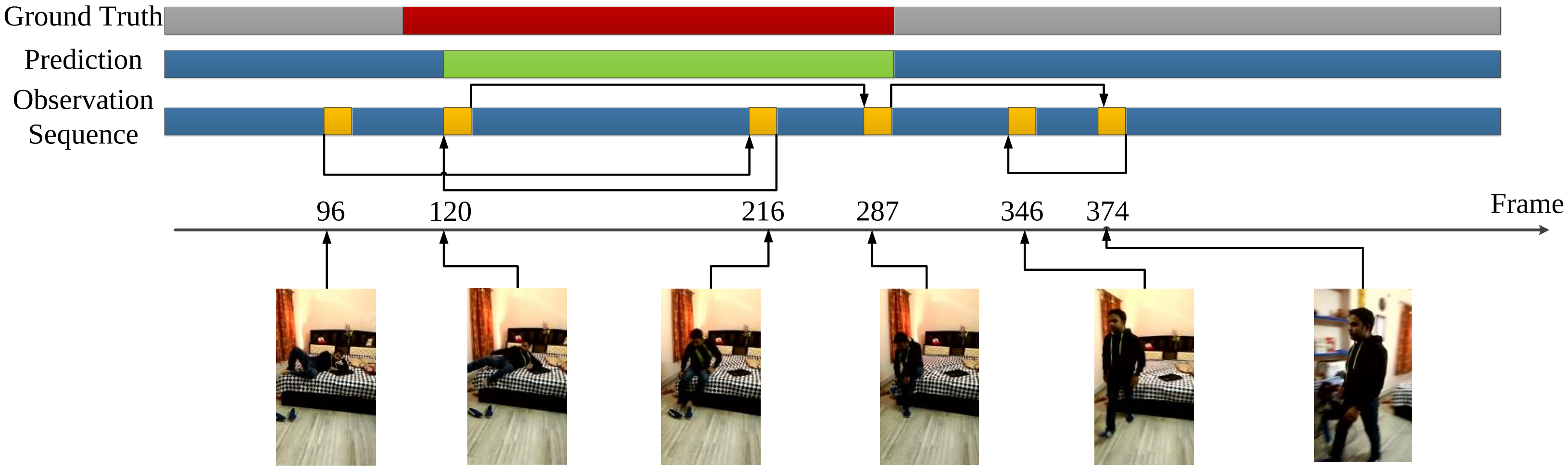
Method	Average running time (per minute video)
CTRL	202ms
Ours	**32ms**

- Simantic concepts lead to significant performance improvement
- Achieve the best performance with 6× faster speed

Example Analysis



Query A: The person washes the leeks in the sink



Query B: Person put on a pair of shoes

Selectively observe a sequence of video frames

The agent can skip in both forward and backward directions in a video

Thank You !