Microsoft

CVPR

微软亚洲研究院创研论坛

CVPR 2020 论文分享会

Weakly-supervised Action Localization by Generative Attention Modeling

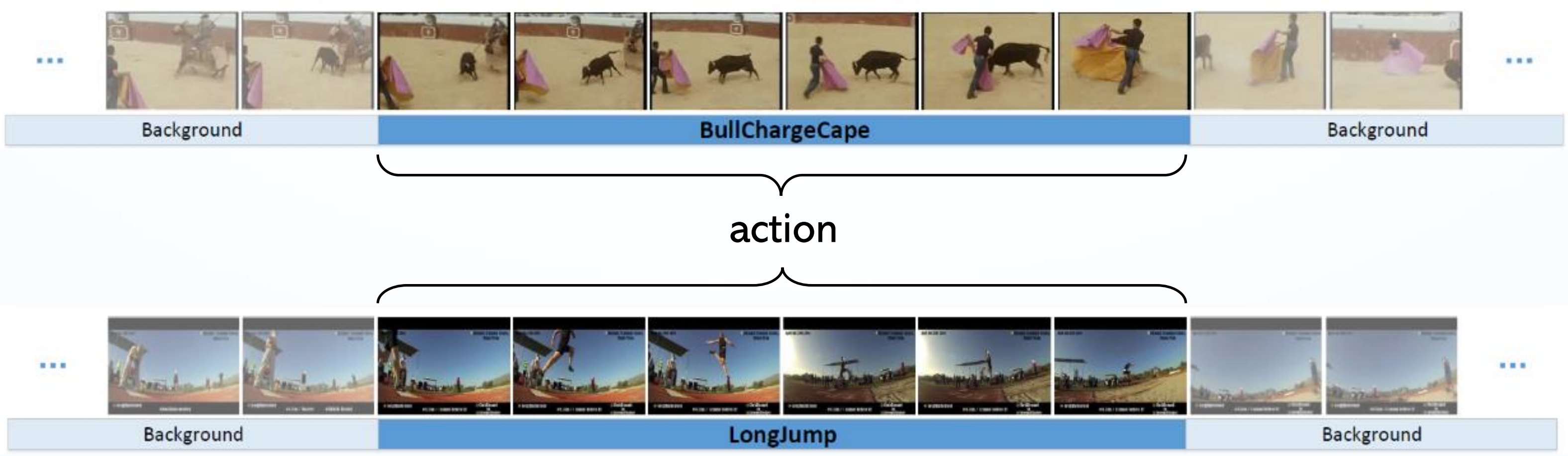
Baifeng Shi[1], **Qi Dai**[2], Yadong Mu[1], Jingdong Wang[2]

[1] Peking University
[2] Microsoft Research Asia

01 Problem of Action Localization

- Action Localization
  - Temporally localize specific actions in an untrimmed video

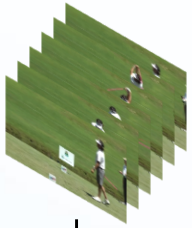
- Examples:



02 Weakly-supervised Action Localization

- Weakly-supervised action localization
  - Save the manual labeling cost
  - Incorporate more training data

Top-down Methods

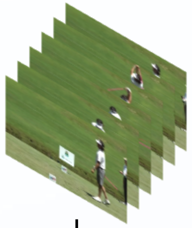


Video-level Classification

Temporal Class Activation Map (TCAM)

[Paul, ECCV'18], [Liu, CVPR'19], [Narayan, ICCV'19]

Bottom-up Methods



Attention

Video-level Classification

[Nguyen, CVPR'18], [Shou, ECCV'18], [Nguyen, ICCV'19]

03 Challenges in Weakly-supervised Action Localization

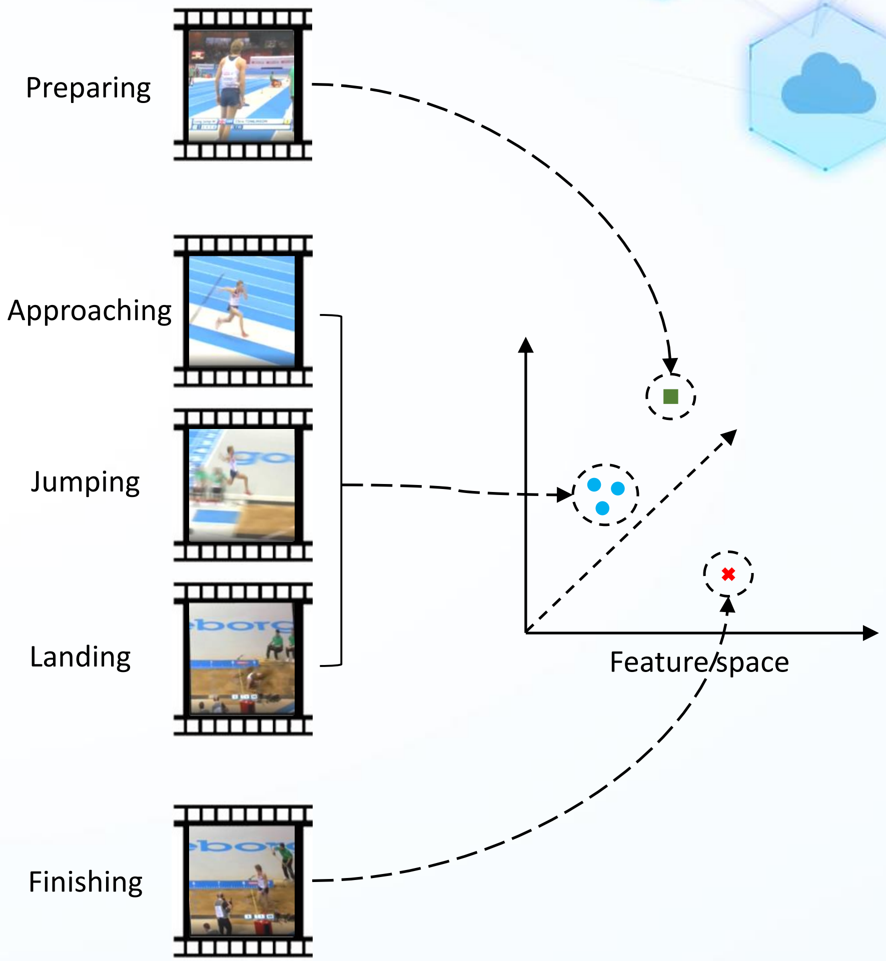
- Classification-based framework

- Existing action-context confusion issue
  - Context: special kind of background with high attention values
  - E.g. track field, sandpit

04 Motivation

- The underlying discrepancy
  - Feature/representation level
  - E.g. more intense body postures

- Solution: Model the representation distribution
  - Different attentions correspond to different features



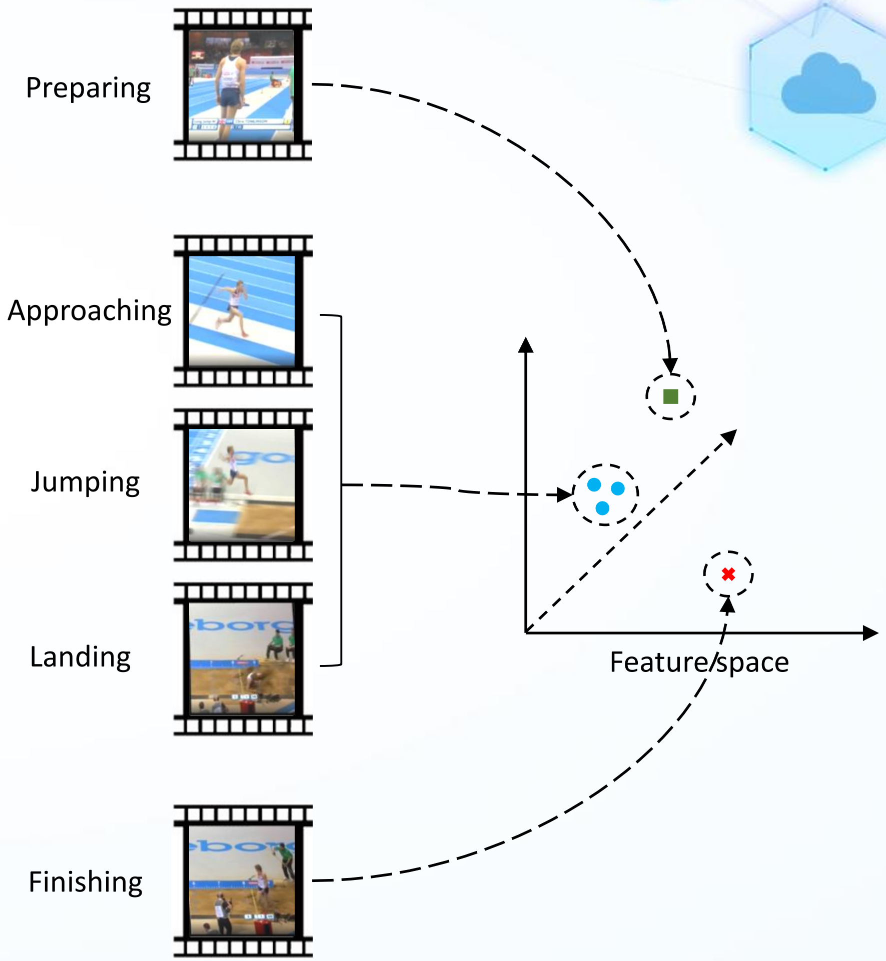
05 Motivation

- Attention-based framework
  - Predict attention λ given video X and label y

$$\max_{\lambda_t \in [0,1]} \log p(\boldsymbol{\lambda}|\mathbf{X}, y)$$

- Optimizing λ involves two aspects
  - High discriminative capacity for classification
  - Accurate prediction of X on λ

$$\begin{aligned}\log p(\boldsymbol{\lambda}|\mathbf{X}, y) &= \log p(\mathbf{X}, y|\boldsymbol{\lambda}) + \log p(\boldsymbol{\lambda}) - \log p(\mathbf{X}, y) \\ &= \log p(y|\mathbf{X}, \boldsymbol{\lambda}) + \log p(\mathbf{X}|\boldsymbol{\lambda}) + \log p(\boldsymbol{\lambda}) \\ &\quad - \log p(\mathbf{X}, y) \\ &\propto \log p(y|\mathbf{X}, \boldsymbol{\lambda}) + \log p(\mathbf{X}|\boldsymbol{\lambda}),\end{aligned}$$



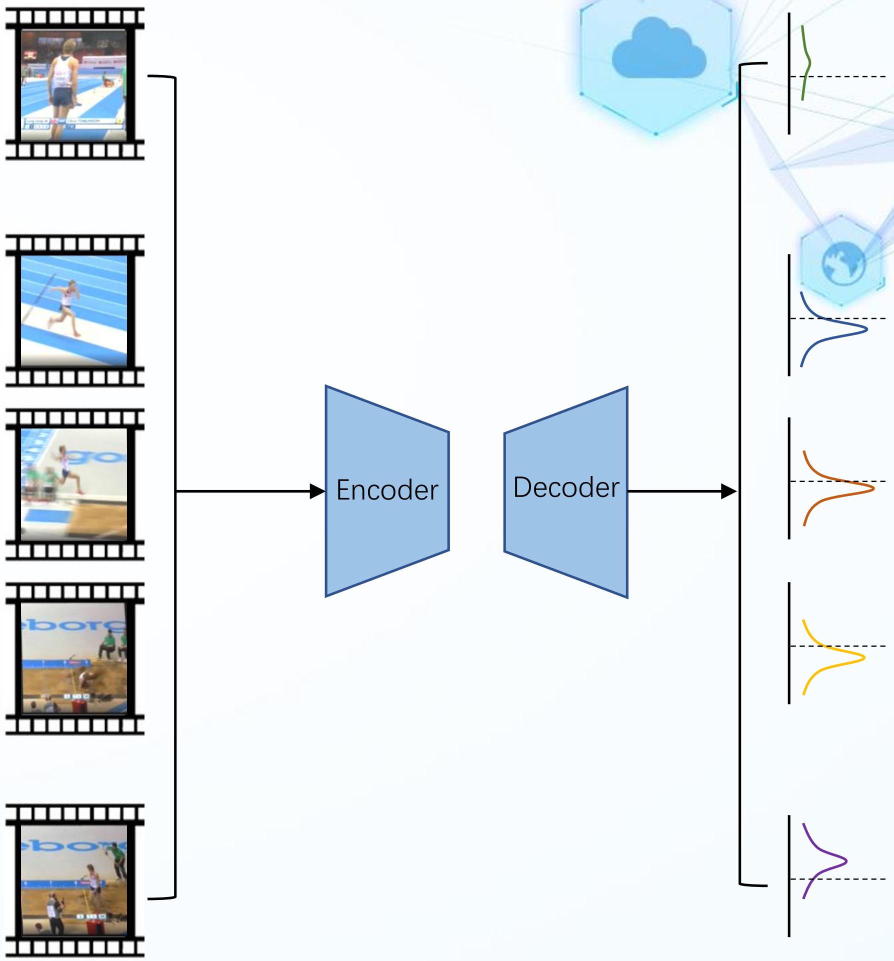
06 Generative Attention Modeling

- Conditional Variational Autoencoder (CVAE)
  - Model the representation distribution conditioned on attention values

$$p_\psi(\mathbf{x}_t|\lambda_t) = \mathbb{E}_{p_\psi(\mathbf{z}_t|\lambda_t)}[p_\psi(\mathbf{x}_t|\lambda_t, \mathbf{z}_t)]$$

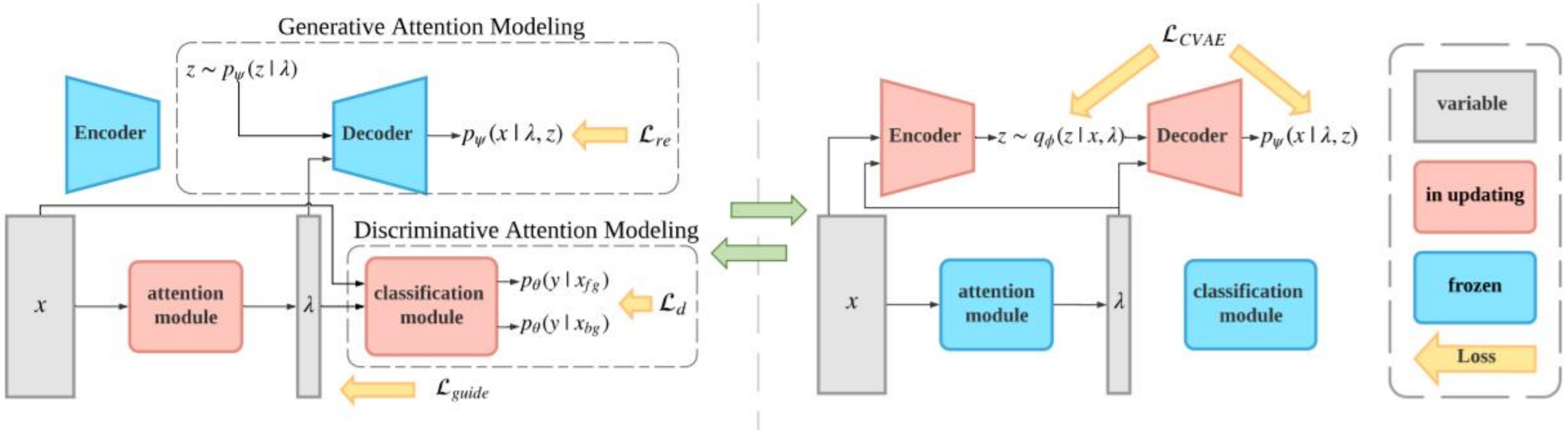
- CVAE can cluster the frames into action, context, and background
  - Frames within each cluster (action, context, or background) can have similar attention values

$$\mathcal{L}_{CVAE} = -\mathbb{E}_{q_\phi(\mathbf{z}_t|\mathbf{x}_t,\lambda_t)} \log p_\psi(\mathbf{x}_t|\lambda_t, \mathbf{z}_t) + \beta \cdot KL(q_\phi(\mathbf{z}_t|\mathbf{x}_t, \lambda_t)||p_\psi(\mathbf{z}_t|\lambda_t))$$



07 Generative Attention Modeling

- Optimize the CVAE under weak supervision
  - Learn model with two alternating stages

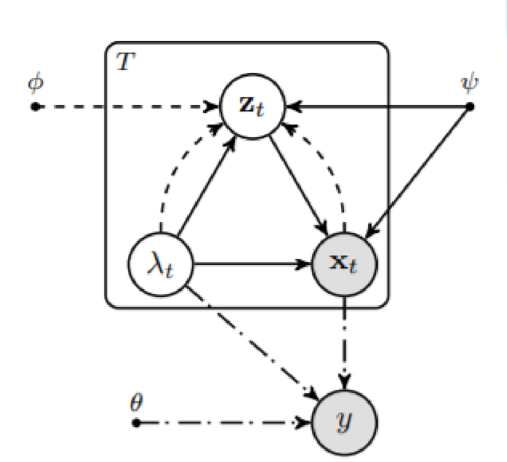


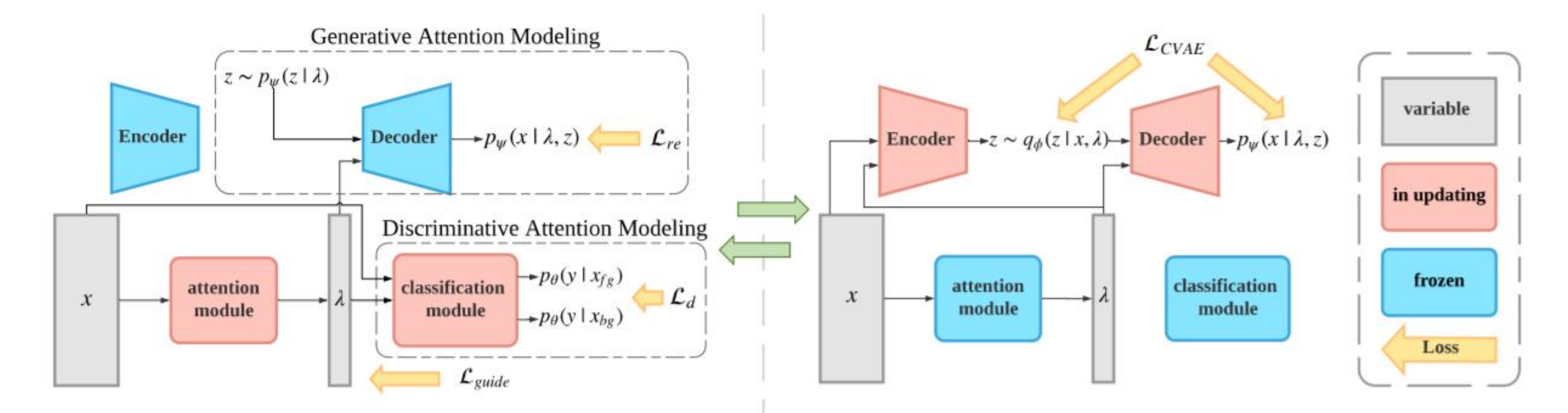
(a) Updating attention / classification modules

(b) Updating CVAE

08 Generative Attention Modeling

- The framework forms a graphical model
  - Generative model ψ
  - Approximation to posterior ϕ
  - Video-level classification model θ





(a) Updating attention / classification modules

(b) Updating CVAE

09 Experiments

- Evaluation of the learned attention values on THUMOS14
  - “Old” model (O): w/o GAM (Generative Attention Model)
  - “New” model (N): w/ GAM
  - Assemble specific models by choosing Attention and Classification modules from O and N

- Attention is the key to achieve good result

Attention	Classification	mAP@IoU				
		0.3	0.4	0.5	0.6	0.7
O	O	43.8	35.8	26.7	18.2	9.7
O	N	44.2	36.1	27.0	18.7	9.8
N	O	46.1	38.2	28.8	19.4	11.2
N	N	46.8	38.2	28.8	19.8	11.4

10 Experiments

- Statistics comparison on THUMOS14
  - "att"/"cls" indicates frame set with high attention/classification scores
  - $|\cdot|$: the size of a set
  - $\uparrow$: higher is better, $\downarrow$: lower is better

Metric		w/o GAM	w/ GAM
$\|att\text{-}gt\| \ / \ \|gt\|$	$\downarrow$	0.777	0.698
$\|gt\text{-}att\| \ / \ \|gt\|$	$\downarrow$	0.858	0.707
$\|(cls - gt) \cap \overline{att}\| \ / \ \|gt\|$	$\uparrow$	1.522	1.543
$\|(att \cap gt)\text{-}cls\| \ / \ \|gt\|$	$\uparrow$	0.001	0.001

11 Experiments

- Results on THUMOS14
  - 2% improvement over the state-of-the-arts on mAP@IoU=0.5

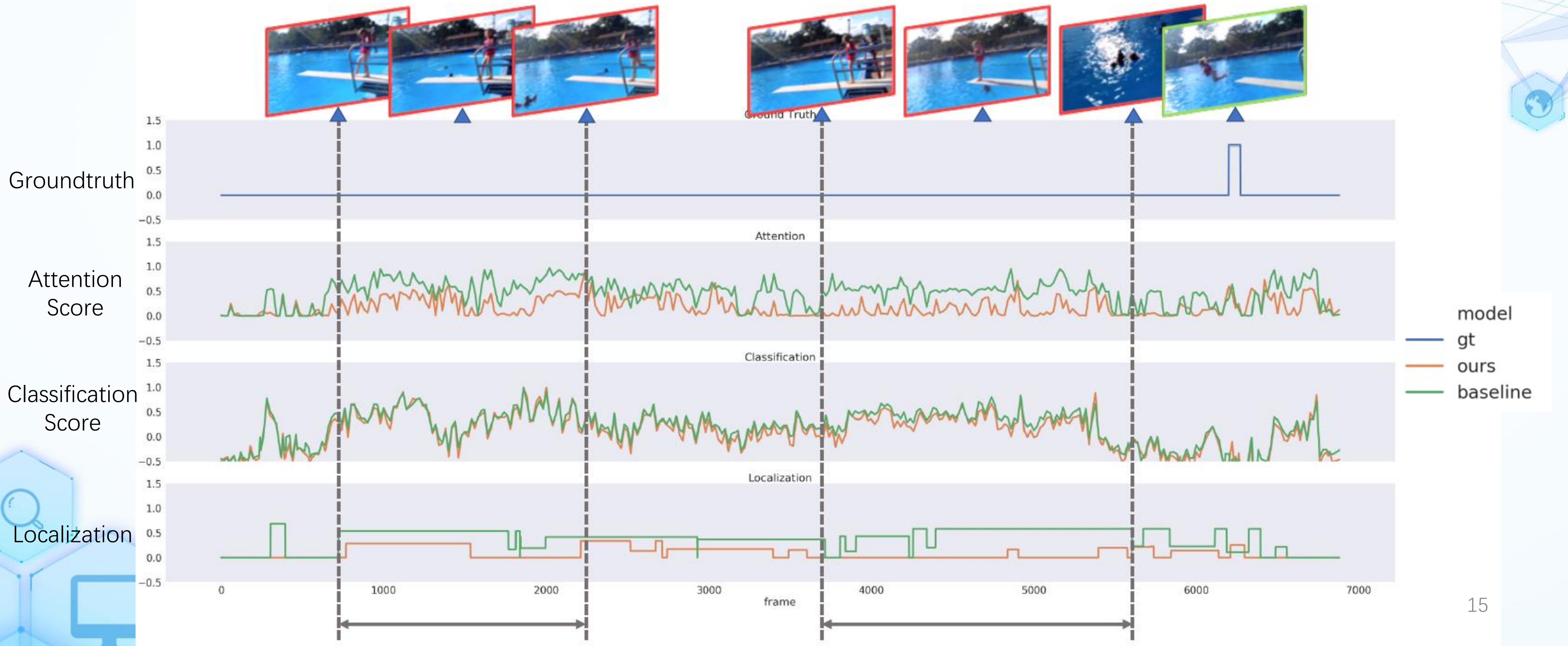
Method	Supervision	Feature	mAP@IoU								
			0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Hide-and-Seek [40]	Weak	-	36.4	27.8	19.5	12.7	6.8	-	-	-	-
UntrimmedNet [45]	Weak	-	44.4	37.7	28.2	21.1	13.7	-	-	-	-
Zhong *et al.* [54]	Weak	-	45.8	39.0	31.1	22.5	15.9	-	-	-	-
AutoLoc [37]	Weak	UNT	-	-	35.8	29.0	21.2	13.4	5.8	-	-
CleanNet [25]	Weak	UNT	-	-	37.0	30.9	23.9	13.9	7.1	-	-
STPN [27]	Weak	I3D	52.0	44.7	35.5	25.8	16.9	9.9	4.3	1.2	0.1
W-TALC [31]	Weak	I3D	55.2	49.6	40.1	31.1	22.8	-	7.6	-	-
Liu *et al.* [23]	Weak	I3D	57.4	50.8	41.2	32.1	23.1	15.0	7.0	-	-
TSM [50]	Weak	I3D	-	-	39.5	-	24.5	-	7.1	-	-
3C-Net [26]	Weak	I3D	56.8	49.8	40.9	32.3	24.6	-	7.7	-	-
Nguyen *et al.* [28]	Weak	I3D	**60.4**	**56.0**	46.6	37.5	26.8	17.6	9.0	3.3	**0.4**
DGAM	Weak	I3D	60.0	54.2	**46.8**	**38.2**	**28.8**	**19.8**	**11.4**	**3.6**	**0.4**

12 Experiments

- Results on ActivityNet1.2
  - 2% improvement over the state-of-the-arts on average mAP

Method	Supervision	mAP@IoU										
		0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95	AVG
SSN [58]	Full	41.3	38.8	35.9	32.9	30.4	27.0	22.2	18.2	13.2	6.1	26.6
UntrimmedNet* [49]	Weak	7.4	6.1	5.2	4.5	3.9	3.2	2.5	1.8	1.2	0.7	3.6
AutoLoc* [41]	Weak	27.3	24.9	22.5	19.9	17.5	15.1	13.0	10.0	6.8	3.3	16.0
W-TALC [34]	Weak	37.0	33.5	30.4	25.7	14.6	12.7	10.0	7.0	4.2	1.5	18.0
TSM [54]	Weak	28.3	26.0	23.6	21.2	18.9	17.0	14.0	11.1	7.5	3.5	17.1
3C-Net [29]	Weak	35.4	-	-	-	22.9	-	-	-	8.5	-	21.1
CleanNet [26]	Weak	37.1	33.4	29.9	26.7	23.4	20.3	17.2	13.9	9.2	5.0	21.6
Liu *et al.* [24]	Weak	36.8	-	-	-	-	22.0	-	-	-	**5.6**	22.4
DGAM	Weak	**41.0**	**37.5**	**33.5**	**30.1**	**26.9**	**23.5**	**19.8**	**15.5**	**10.8**	5.3	**24.4**

13 Experiments



14 Reference

[1] Paul S, Roy S, Roy-Chowdhury A K. W-talc: Weakly-supervised temporal activity localization and classification. ECCV, 2018: 563-579.
[2] Liu D, Jiang T, Wang Y. Completeness modeling and context separation for weakly supervised temporal action localization. CVPR, 2019: 1298-1307.
[3] Narayan S, Cholakkal H, Khan F S, et al. 3c-net: Category count and center loss for weakly-supervised action localization. ICCV, 2019: 8679-8687.
[4] Shou Z, Gao H, Zhang L, et al. Autoloc: Weakly-supervised temporal action localization in untrimmed videos. ECCV, 2018: 154-171.
[5] Nguyen P, Liu T, Prasad G, et al. Weakly supervised action localization by sparse temporal pooling network. CVPR, 2018: 6752-6761.
[6] Nguyen P X, Ramanan D, Fowlkes C C. Weakly-supervised action localization with background modeling. ICCV, 2019: 5502-5511.

Code Available:
https://github.com/bfshi/DGAM-Weakly-Supervised-Action-Localization

谢谢观看

THANK YOU