UTS

Towards Efficient and Effective Video Analysis

Yi Yang

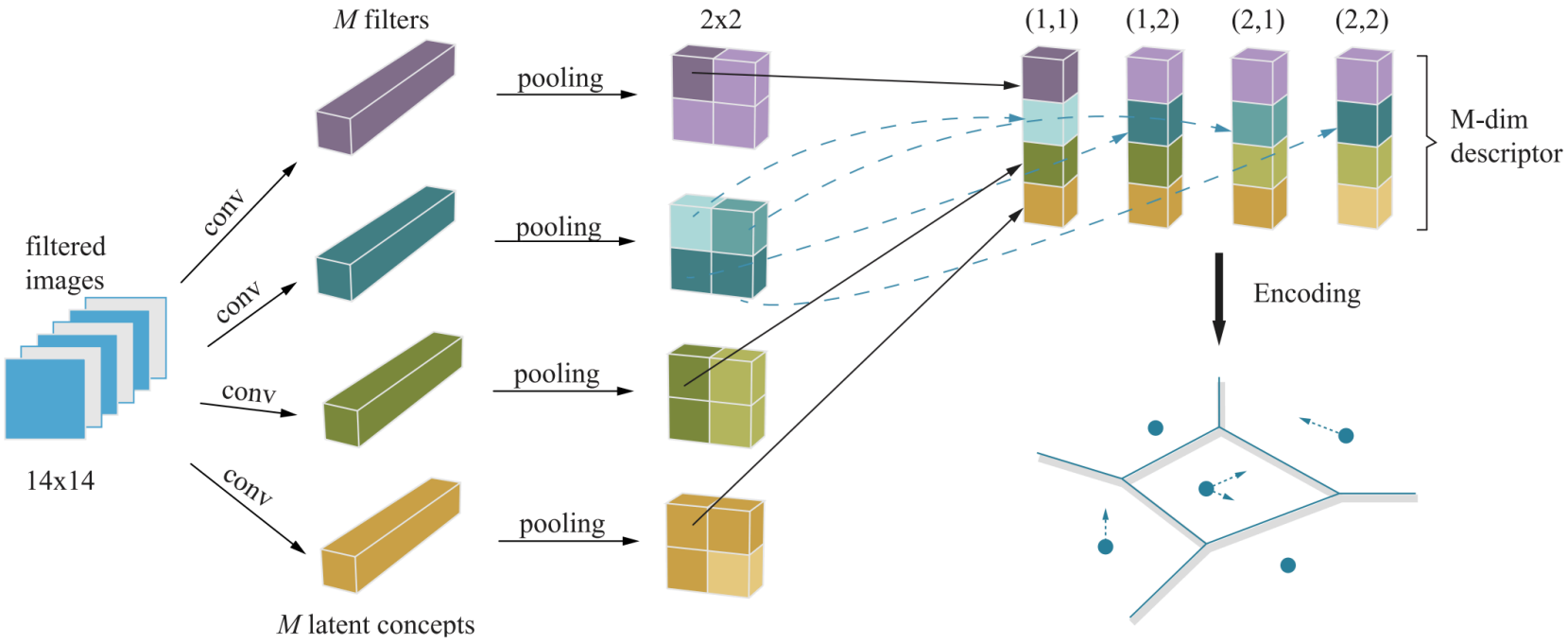
ReLER Recognition, LEarning, Reasoning

Overview

- Third-person action analysis
  - Untrimmed video classification
  - Video redundancy for efficient video analysis
  - Annotation-efficient action localization
- Effective ego-centric video analysis
  - Winning solution at EPIC-KITCHENS Action Recognition Competition 2019, 2020
- Video object segmentation

Untrimmed video classification

Feature Aggregation for ConvNets



Vector of Locally Aggregated Descriptors (VLAD) encoding

$$\boldsymbol{u}_k = \sum_{i:\mathrm{NN}(\boldsymbol{x}_i)=\boldsymbol{c}_k} (\boldsymbol{x}_i - \boldsymbol{c}_k)$$

Zhongwen Xu, Yi Yang, Alex G Hauptmann, A discriminative CNN video representation for event detection, CVPR 2015.

Feature Aggregation for ConvNets

- Ranked 1st in THUMOS 2015

THUMOS Challenge (2015) Results:

Action Classification Task:

Rank	Entry	Run1	Run2	Run3	Run4	Run5
1	U. of Tech., Sydney & CMU	0.7384	0.7157	0.7011	0.6913	0.647
2	MSR Asia (MSM)	0.6861	0.6869	0.6878	0.6886	0.6897
3	Zhejiang University	0.6876	0.6643	0.6859	0.6809	0.5625
4	INRIA_LEAR	0.6814	0.6811	0.5395	0.6739	0.6793
5	CUHK & Shenzhen Inst. Adv. Tech.	0.4894	0.5746	0.6803	0.6576	0.6604
6	University of Amsterdam	0.6798	NA	NA	NA	NA

5%+ gains

Multi-rate Modeling

- High frame rate contains many redundant frames, which would require more computation to process;
- Low frame rate may miss important frames;
- Leverage video multi-rate encoding with different intervals.

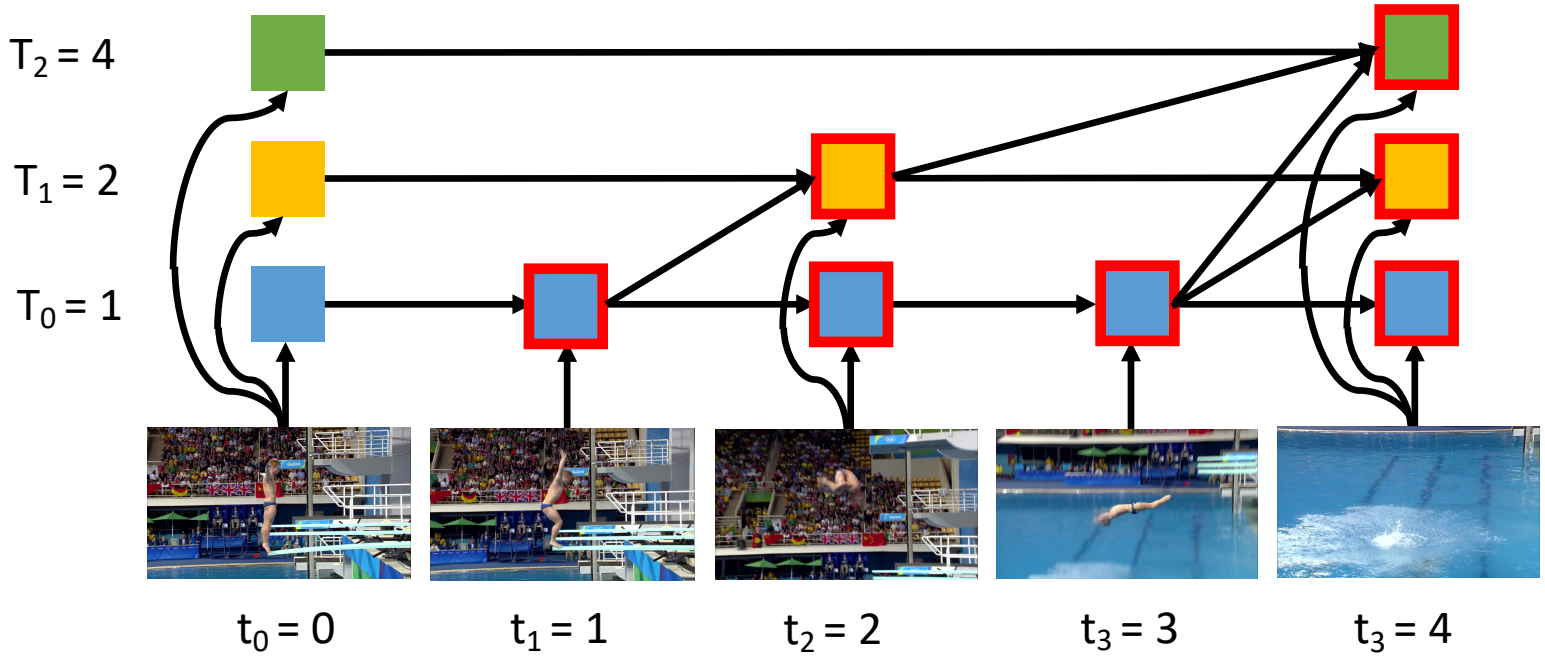
30 FPS

2 FPS

2 FPS (Delay 1s)

Multi-rate Modeling

- Sparse frames extract slow motion information
- Dense frames extract fast motion
- Leverage both slow and fast motion information for video representation learning



Linchao Zhu, Zhongwen Xu, Yi Yang, Bidirectional Multirate Reconstruction for Temporal Modeling in Videos, CVPR 2017

Multi-rate Modeling

- Evaluation on Multimedia Event Detection. State-of-the-art was achieved.

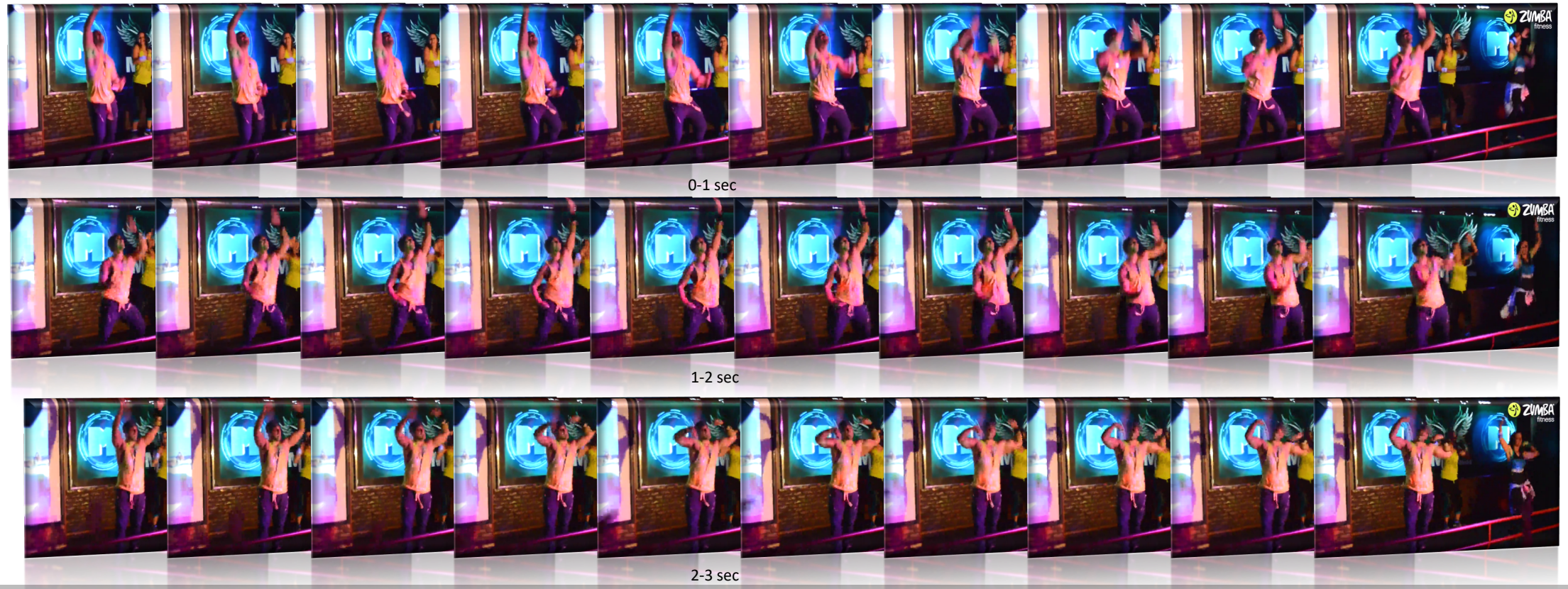
Model	MEDTest-13	MEDTest-14
C3D	36.9	31.4
VGG16+LCD+VLAD	40.3	35.7
Ours	44.5 (+4.2%)	37.3 (+1.6%)

Linchao Zhu, Zhongwen Xu, Yi Yang, Bidirectional Multirate Reconstruction for Temporal Modeling in Videos, CVPR 2017

Efficient video analysis

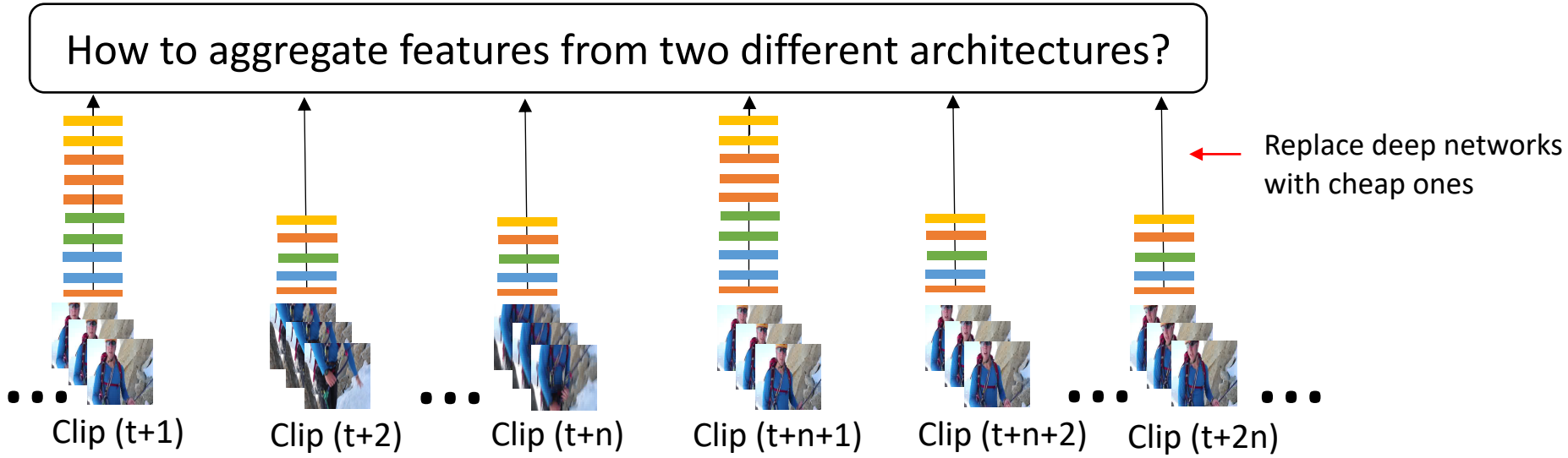
Efficient video classification

- Video frames are redundant.



Efficient video classification

- Adjacent clips are redundant
- The idea is to replace deep networks with cheap ones

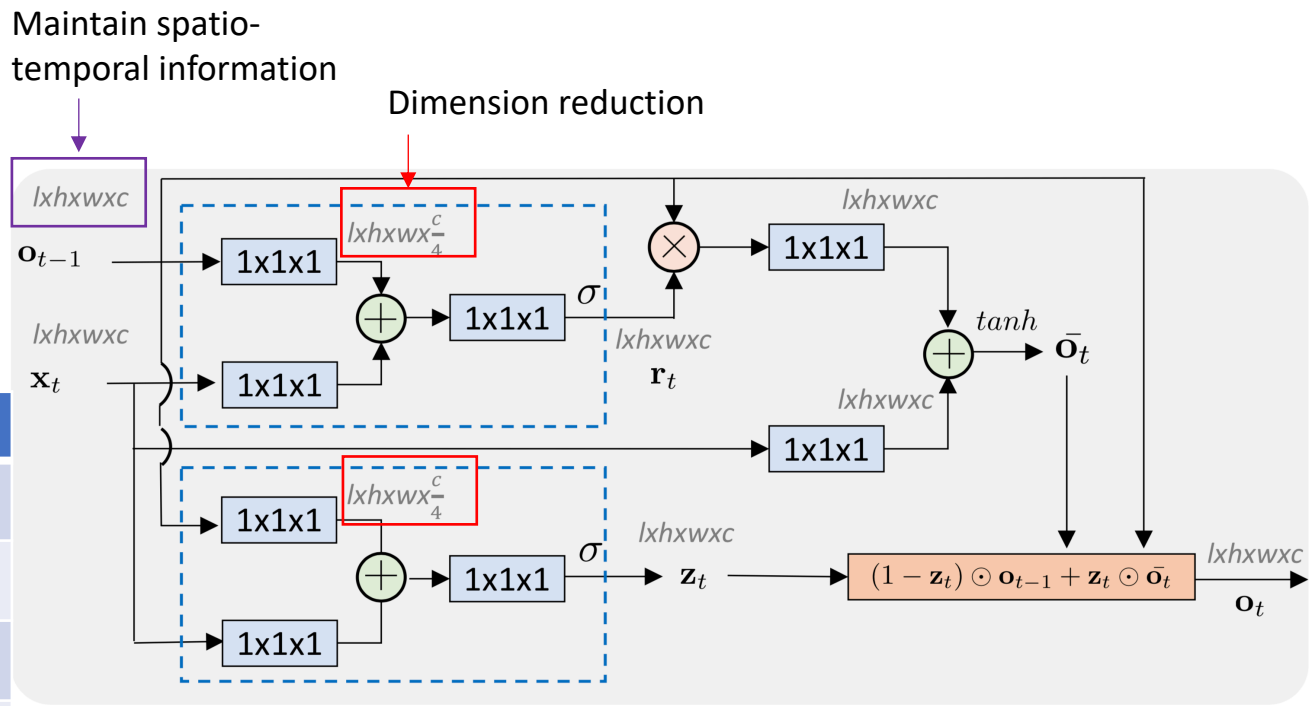


FASTER Recurrent Networks for Efficient Video Classification, AAAI 2020

Efficient video classification

- Introduce Fast-GRU to integrate mixed networks
  - ➢ Maintain spatio-temporal information. Beneficial to long-term sequence modeling
  - ➢ Improve the discriminative ability of the gate (input gate, update gate, etc.) by introducing more nonlinearity

Model	#clip=2	#clip=4	#clip=8	#clip=16	#clip=32
Avg. pool	60.7	63.9	64.2	64.0	63.8
Concat	61.6	65.9	66.3	62.3	58.2
LSTM	61.3	66.2	66.3	66.3	64.7
GRU	61.1	65.6	66.5	66.1	65.8
FAST-GRU	61.1	65.6	**67.2**	**67.4**	**67.2**



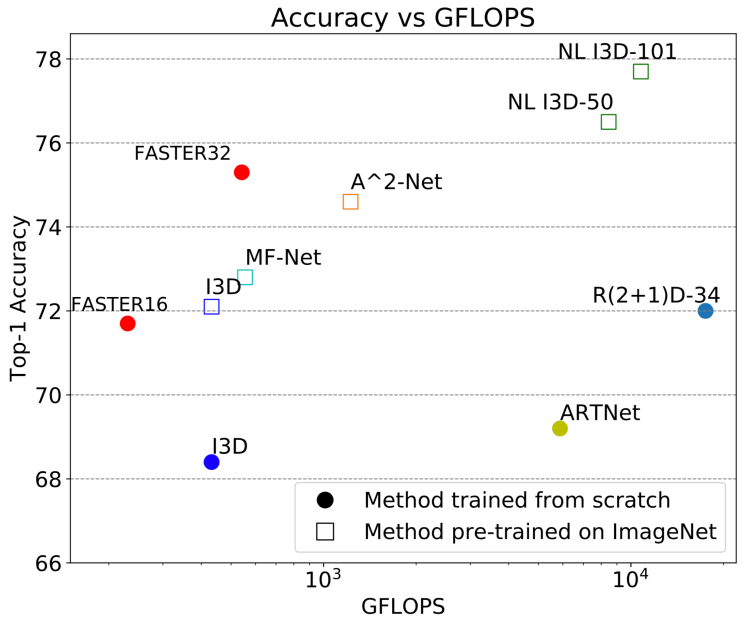
FAST-GRU

Efficient video classification

- Improve 3D networks, e.g., ResNet (2+1)D-34， reducing GFLOPS more than 10 times.

Model	Top-1	ImageNet pre-train	GFLOPs×clips
I3D (Carreira and Zisserman 2017)	72.1	✓	108×4
S3D (Xie et al. 2018)	72.2	✓	66.4×N/A
MF-Net (Chen et al. 2018b)	72.8	✓	11.1×50
A^2-Net (Chen et al. 2018a)	74.6	✓	40.8×30
S3D-G (Xie et al. 2018)	74.7	✓	71.4×N/A
NL I3D-50 (Wang et al. 2018)	76.5	✓	282×30
NL I3D-101 (Wang et al. 2018)	77.7	✓	359×30
I3D (Carreira and Zisserman 2017)	68.4	-	108×4
STC (Diba et al. 2018)	68.7	-	N/A×N/A
ARTNet (Wang et al. 2017)	69.2	-	23.5×250
S3D (Xie et al. 2018)	69.4	-	66.4×N/A
ECO (Zolfaghari, Singh, and Brox 2018)	70.0	-	N/A×N/A
R(2+1)D-34 (Tran et al. 2018)	72.0	-	152×115
FASTER16	71.7	-	14.4×16
FASTER32	**75.3**	-	67.7×8

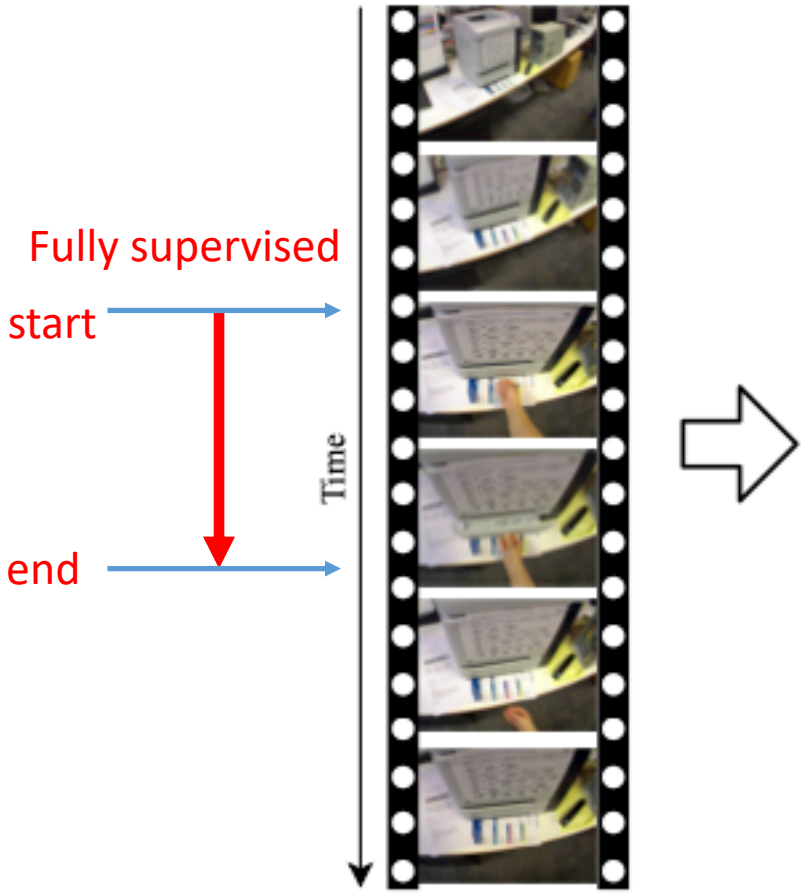
Comparisons on Kinetics

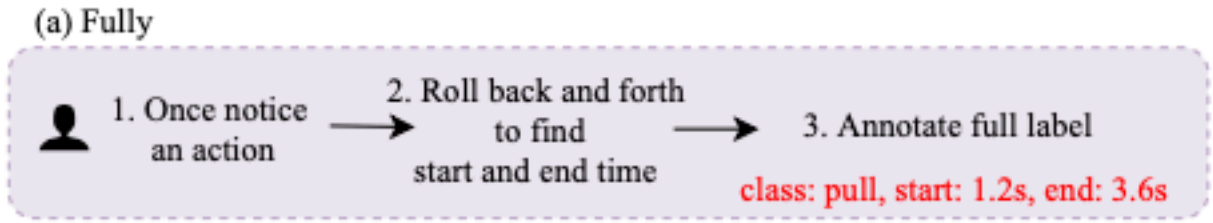


Log-scale GFLOPs *vs*. accuracy comparisons on Kinetics

Annotation-efficient action localization

Annotation-efficient temporal action localization

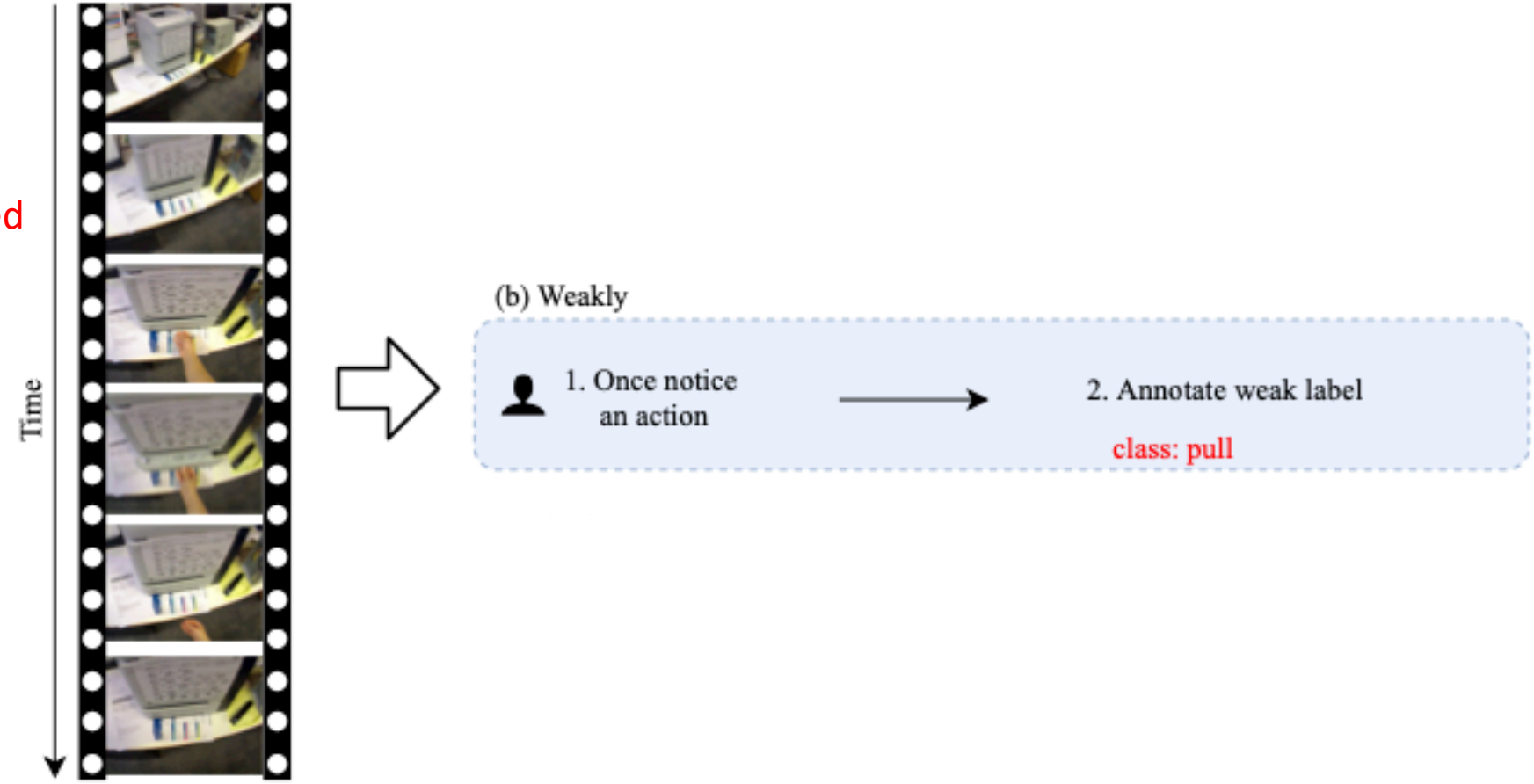




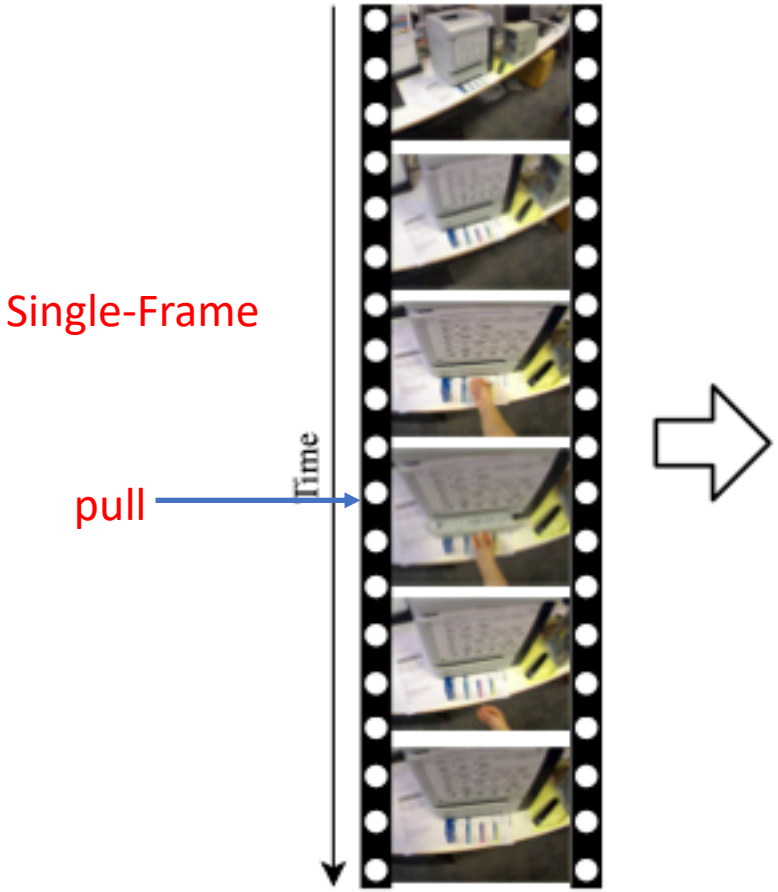
Annotation-efficient temporal action localization

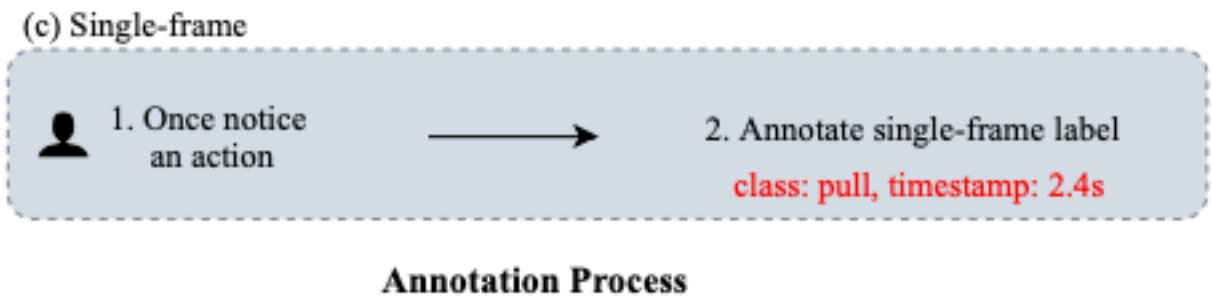
Weakly supervised

pull

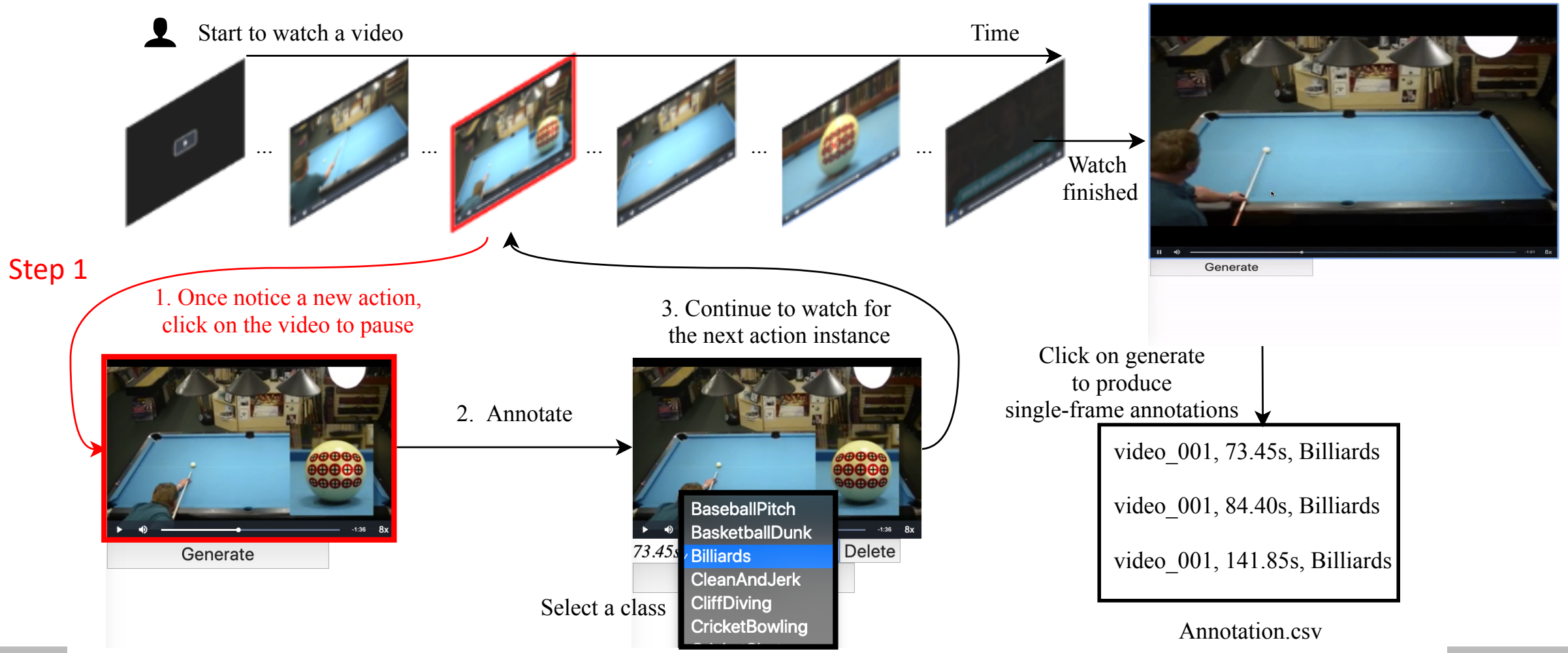


Annotation-efficient temporal action localization

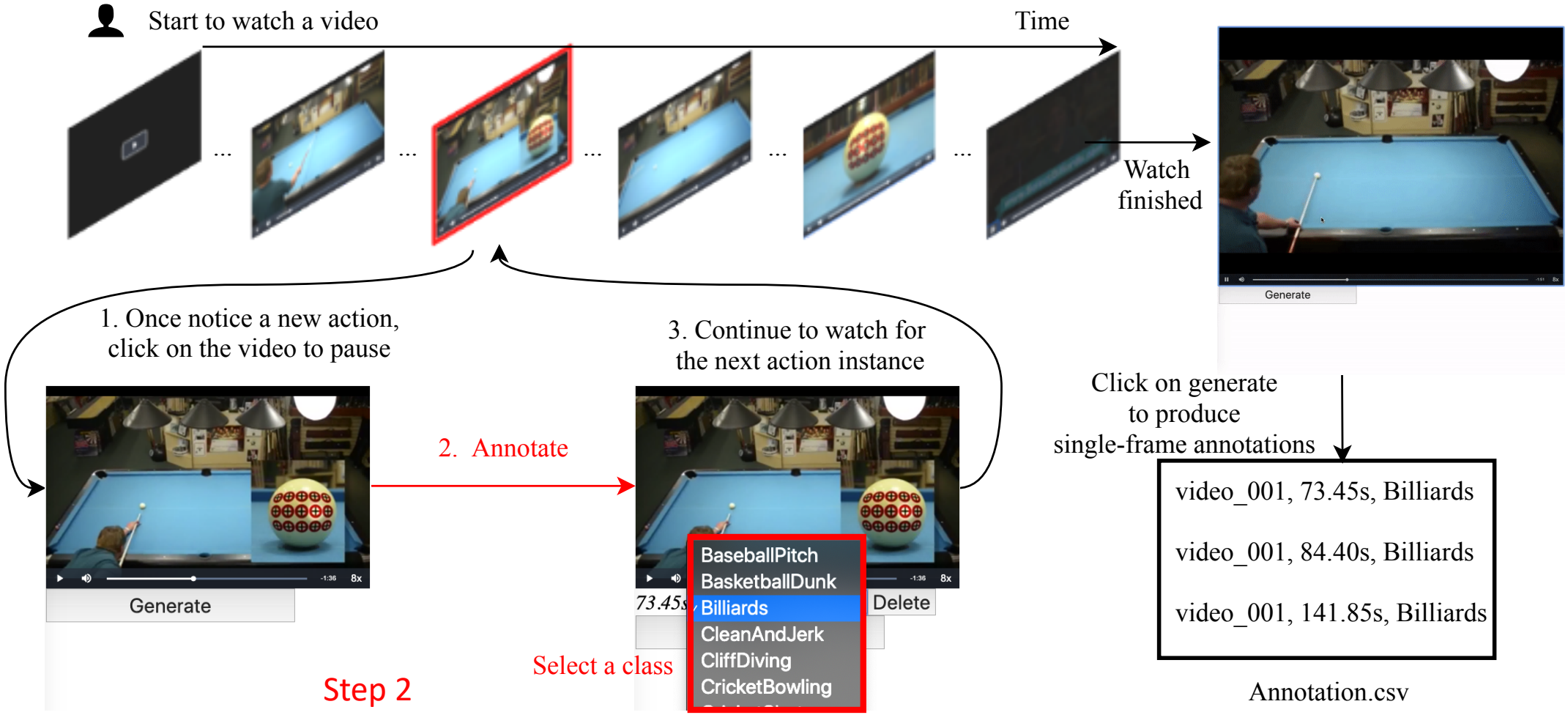




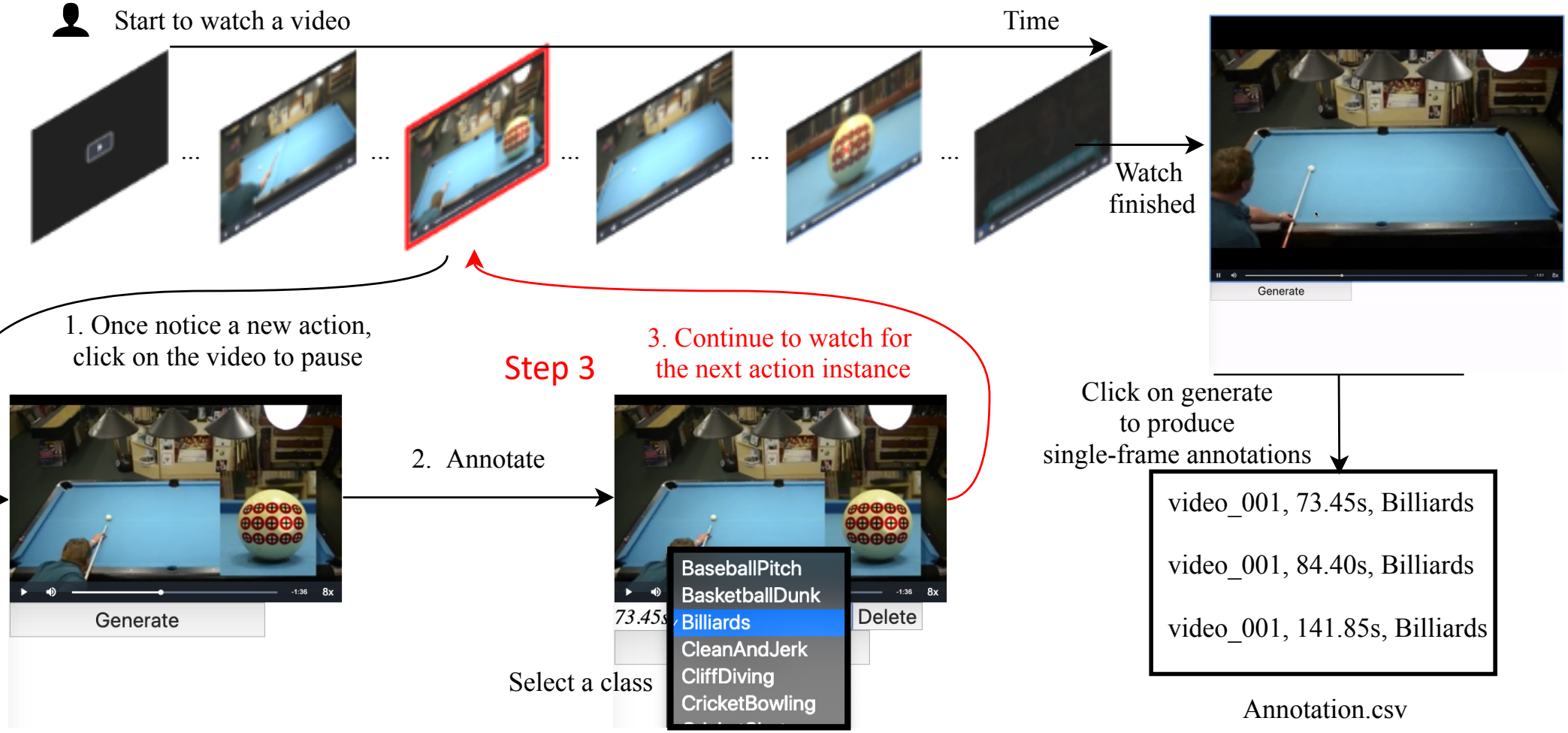
Annotation-efficient temporal action localization



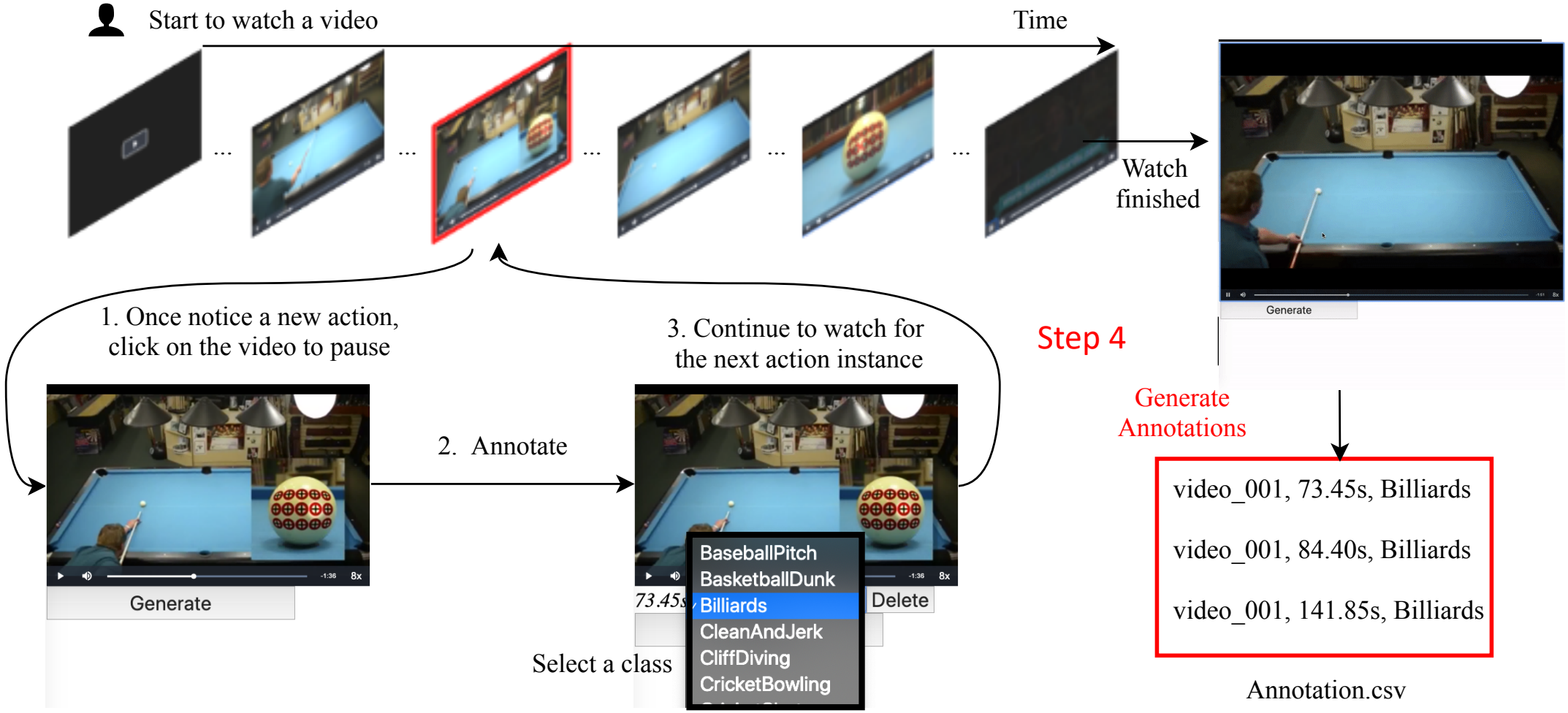
Annotation-efficient temporal action localization



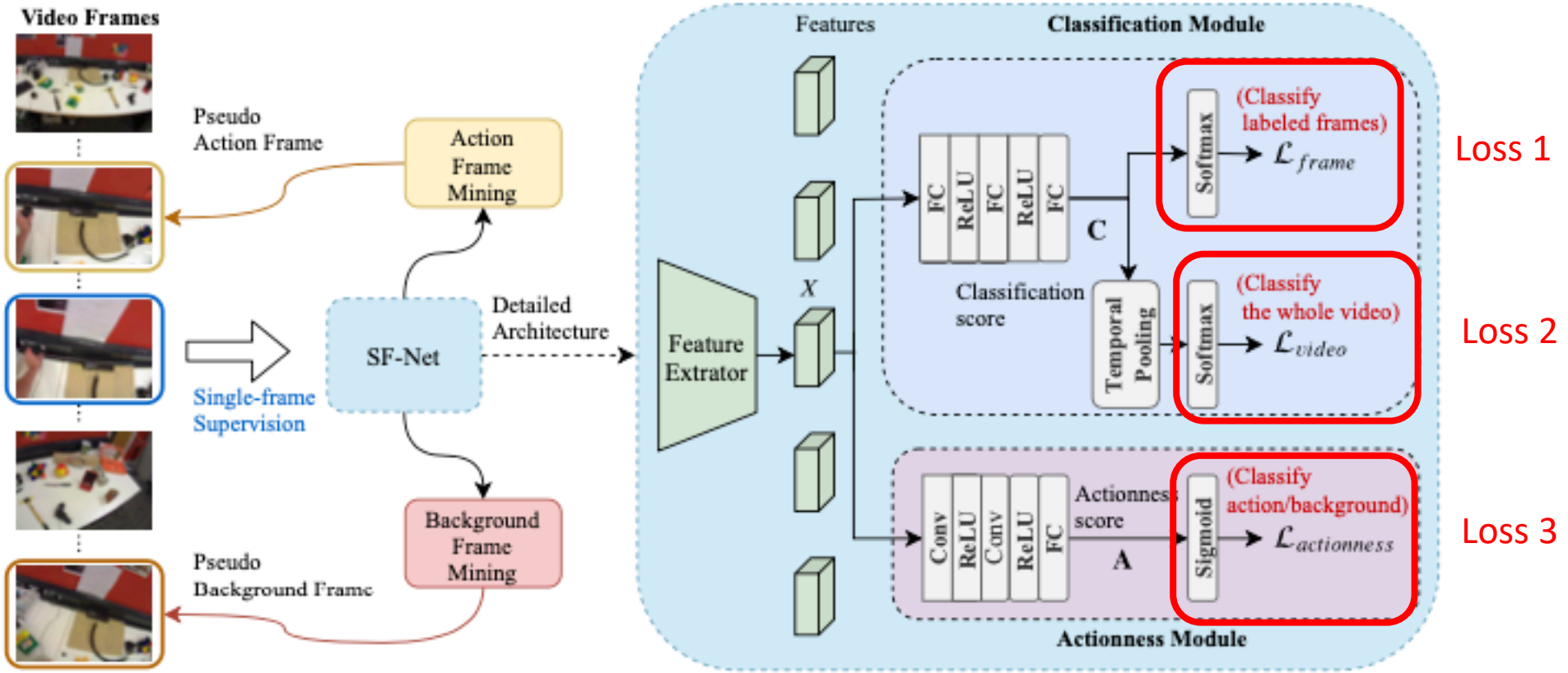
Annotation-efficient temporal action localization



Annotation-efficient temporal action localization



Annotation-efficient temporal action localization



SF-Net: Single-Frame Supervision for Temporal Action Localization, ECCV 2020 (Spotlight)

Annotation-efficient temporal action localization

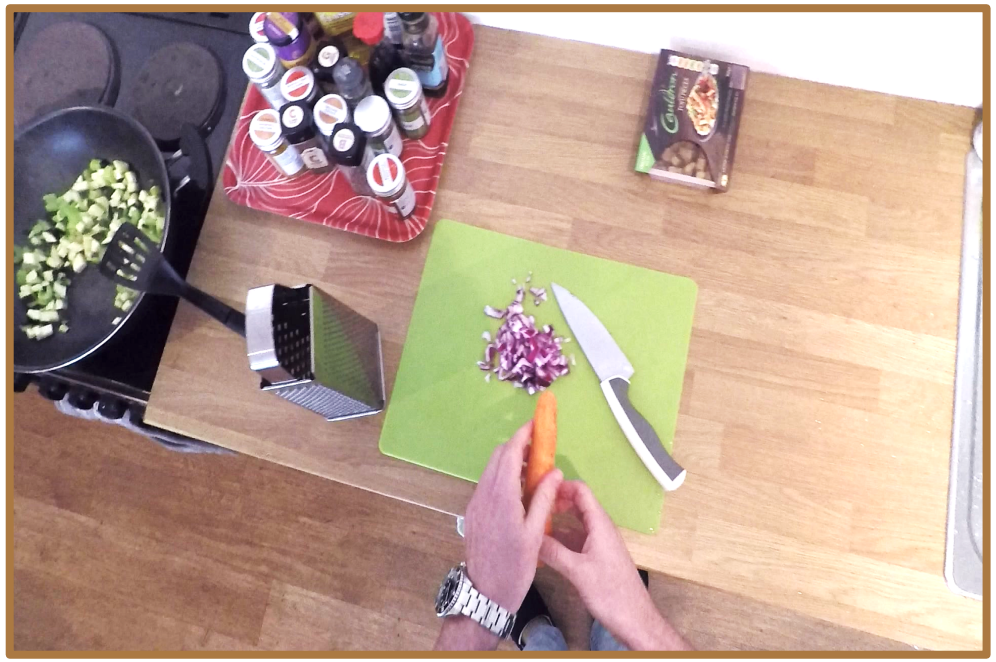
Dataset	Models	mAP@IoU				
		0.1	0.3	0.5	0.7	AVG
GTEA	Full	58.1	40.0	22.2	14.8	31.5
	Weak	14.0	9.7	4.0	3.4	7.0
	SF	50.0±1.42	35.6±2.61	**21.6**±1.67	**17.7**±0.96	30.5±1.23
	SFB	52.9±3.84	34.9±4.72	17.2±3.46	11.0±2.52	28.0±3.53
	SFBA	52.6±5.32	32.7±3.07	15.3±3.63	8.5±1.95	26.4±3.61
	SFBAE	**58.0**±2.83	**37.9**±3.18	19.3±1.03	11.9±3.89	**31.0**±1.63
BEOID	Full	65.1	38.6	22.9	7.9	33.6
	Weak	22.5	11.8	1.4	0.3	8.7
	SF	54.1±2.48	24.1±2.37	6.7±1.72	1.5±0.84	19.7±1.25
	SFB	57.2±3.21	26.8±1.77	9.3±1.94	1.7±0.68	21.7±1.43
	SFBA	**62.9**±1.68	36.1±3.17	12.2±3.15	2.2±2.07	27.1±1.44
	SFBAE	**62.9**±1.39	**40.6**±1.8	**16.7**±3.56	**3.5**±0.25	**30.1**±1.22
THUMOS14	Full	68.7	54.5	34.4	16.7	43.8
	Weak	55.3	40.4	20.4	7.3	30.8
	SF	58.6±0.56	41.3±0.62	20.4±0.55	6.9±0.33	31.7±0.41
	SFB	60.8±0.65	44.5±0.37	22.9±0.38	7.8±0.46	33.9±0.31
	SFBA	68.7±0.33	52.3±1.21	28.2±0.42	**9.7**±0.51	39.9±0.43
	SFBAE	**70.0**±0.64	**53.3**±0.3	**28.8**±0.57	**9.7**±0.35	**40.6**±0.40

Annotation-efficient temporal action localization

Supervision	Method	mAP @IoU							
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	AVG
Full	S-CNN [30]	47.7	43.5	36.3	28.7	19.0	-	5.3	35.0
Full	CDC [28]	-	-	40.1	29.4	23.3	-	7.9	-
Full	R-C3D [37]	54.5	51.5	44.8	35.6	28.9	-	-	43.1
Full	SSN [42]	60.3	56.2	50.6	40.8	29.1	-	-	47.4
Full	Faster- [6]	59.8	57.1	53.2	48.5	42.8	**33.8**	**20.8**	52.3
Full	BMN [16]	-	-	56.0	47.4	38.8	29.7	20.5	-
Full	P-GCN [40]	**69.5**	**67.8**	**63.6**	**57.8**	**49.1**	-	-	**61.6**
Weak	Hide-and-Seek [32]	36.4	27.8	19.5	12.7	6.8	-	-	20.6
Weak	UntrimmedNet [35]	44.4	37.7	28.2	21.1	13.7	-	-	29.0
Weak	W-TALC [10]	49.0	42.8	32.0	26.0	18.8	-	6.2	33.7
Weak	AutoLoc [29]	-	-	35.8	29.0	21.2	13.4	5.8	-
Weak	STPN [25]	52.0	44.7	35.5	25.8	16.9	9.9	4.3	35.0
Weak	W-TALC [27]	55.2	49.6	40.1	31.1	22.8	-	7.6	39.7
Weak	Liu *et al.* [19]	57.4	50.8	41.2	32.1	23.1	15.0	7.0	40.9
Weak	Nguyen *et al.* [26]	**60.4**	**56.0**	**46.6**	**37.5**	**26.8**	**17.6**	**9.0**	**45.5**
Weak	3C-Net [24]	59.1	53.5	44.2	34.1	26.6	-	8.1	43.5
Single-frame simulation*	Moltisanti *et al.* [23]	24.3	19.9	15.9	12.5	9.0	-	-	16.3
Single-frame simulation*	SF-Net	68.3	62.3	52.8	**42.2**	**30.5**	**20.6**	**12.0**	51.2
Single-frame#	SF-Net	**71.0**	**63.4**	**53.2**	40.7	29.3	18.4	9.6	**51.5**

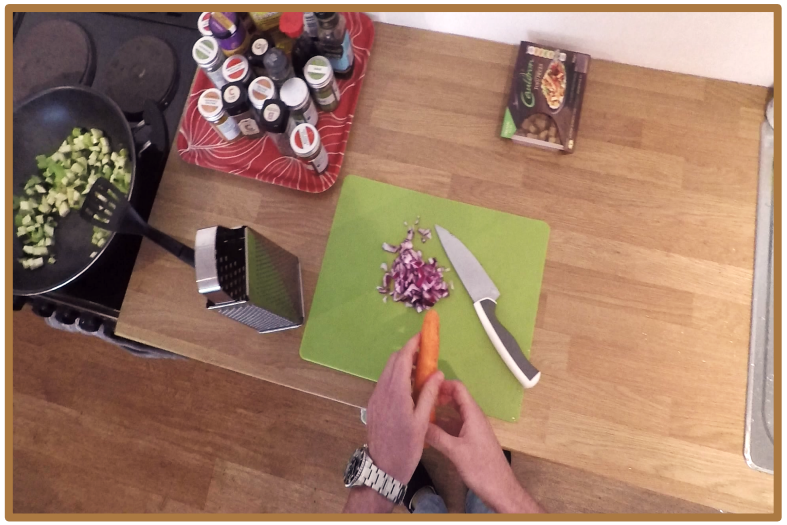
Ego-centric action recognition

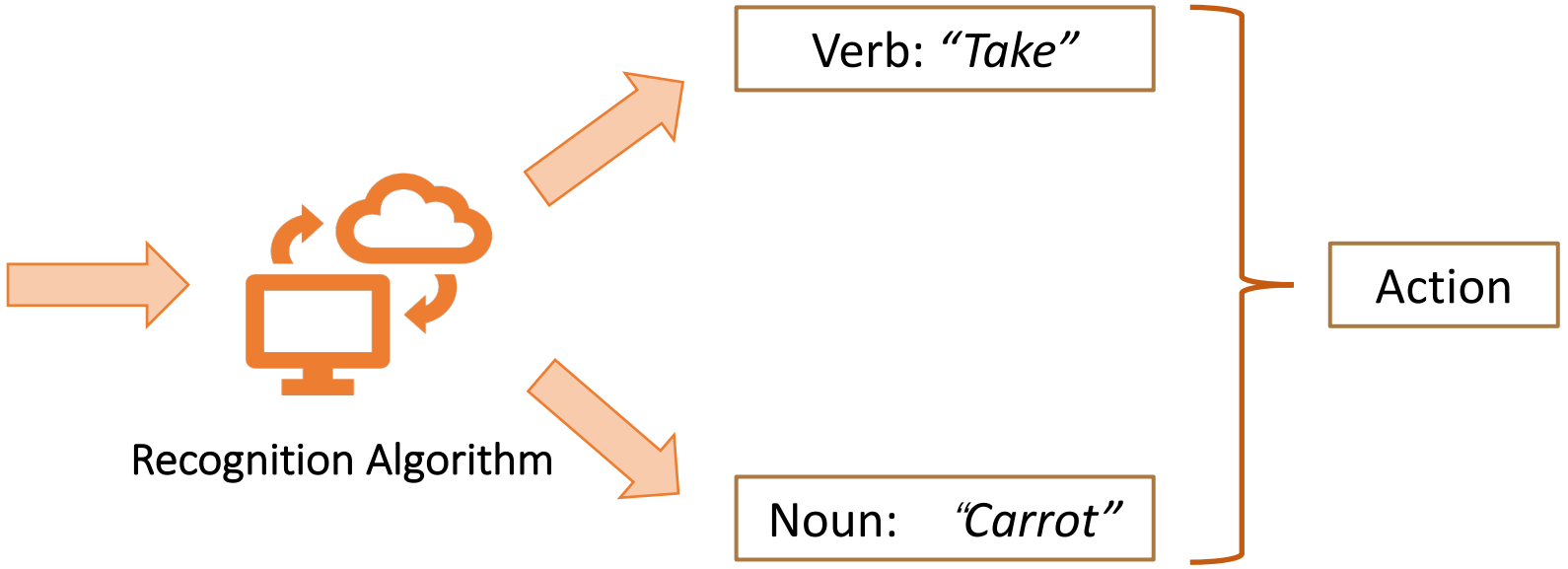
Ego-centric videos



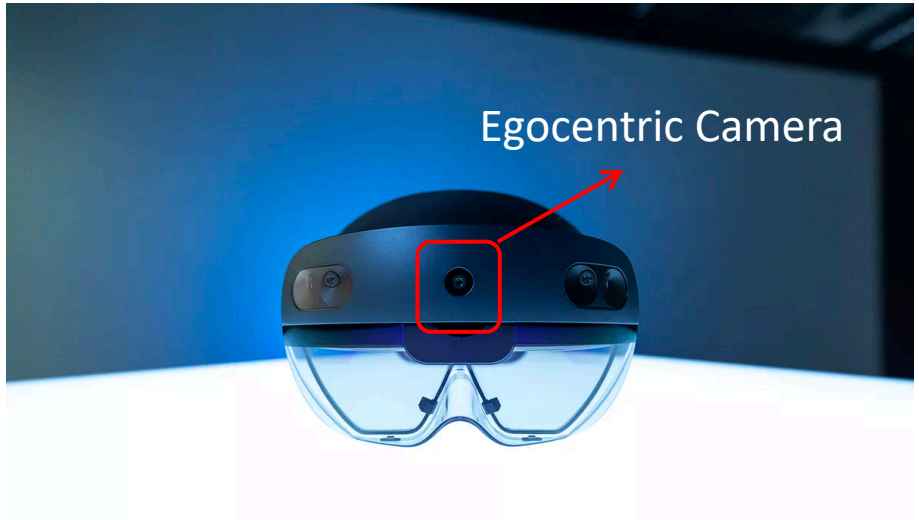
Damen et al. *Scaling egocentric vision: the EPIC-KITCHENS dataset.* In ECCV, 2018

Egocentric Action Recognition





Applications



Virtual Reality / Augmented Reality

Human-to-robot imitation Learning

[1] https://www.theverge.com/2019/11/7/20946589/microsoft-hololens-2-mixed-reality-headset-preorder-shipping-price-upgrade

[2] Zhang et al. *Deep imitation learning for complex manipulation tasks from virtual reality.* In ICRA, 2018

Compare to Third-person Videos

Third-person Action Recognition

- ✓ Human-centric
- ✓ Scene-relevant
- ✓ Coarse-grained

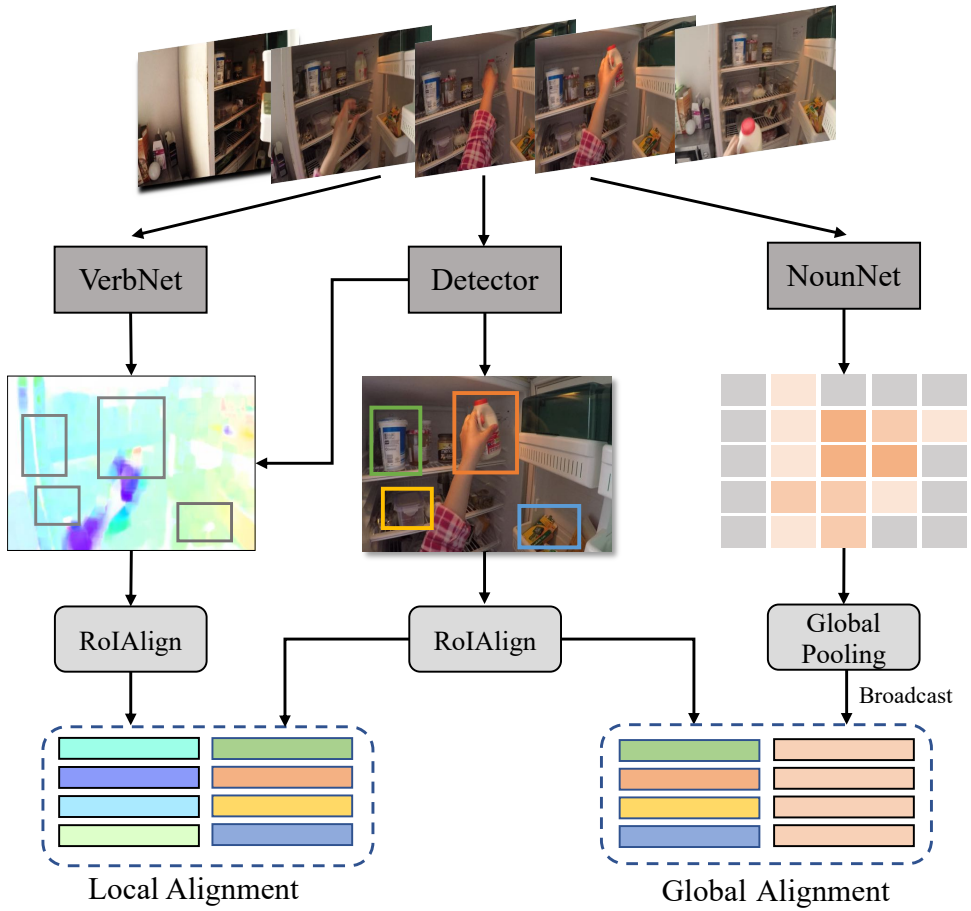
First-person Action Recognition

- ✓ Object-centric
- ✓ Scene-irrelevant
- ✓ Fine-grained

Challenges

- Objects are small and various in videos
- large camera motion and confusing interaction locations

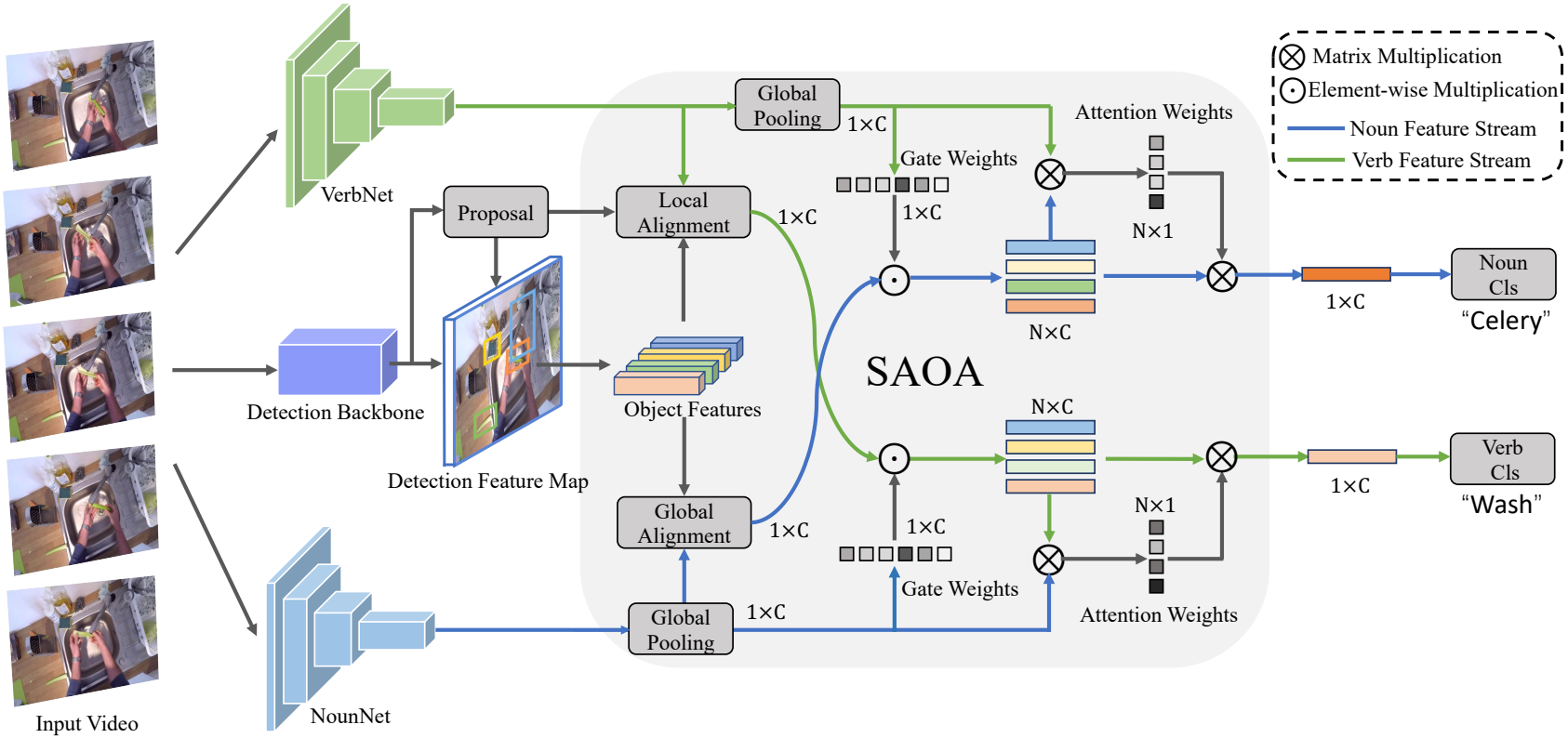
Our Method



Object-centric feature alignment

- Global alignment for noun classification.
- Local alignment for verb classification.

Symbiotic Attention



- Object detection offers detailed local understanding
- Symbiotic attention enables mutual interactions and meticulous reasoning

Symbiotic Attention for Egocentric Action Recognition with Object-centric Alignment, TPAMI 2020

Datasets

Epic-Kitchens is the **largest** video dataset in first-person vision.

- 55 hours of recording (Full HD, 60fps)
- 39,594 action segments
- 125 verb classes, 331 noun classes

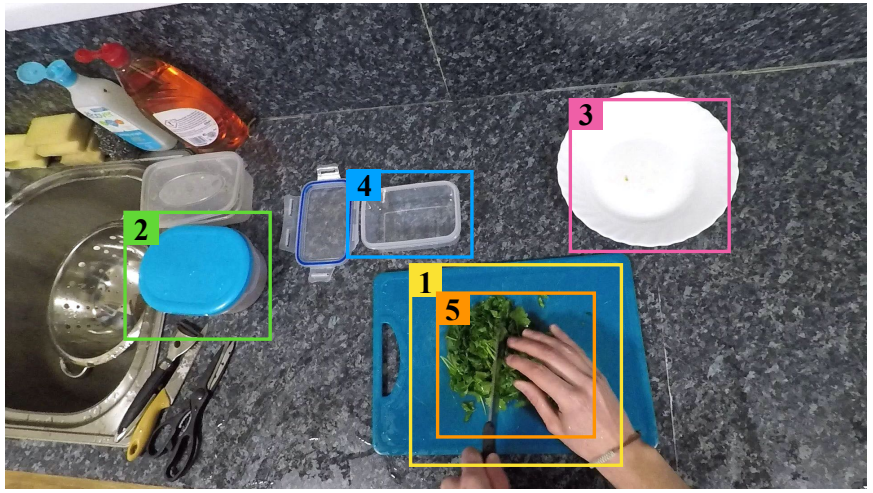
Evaluation results

Table 3. Results on the EPIC-Kitchens leaderboard.

	Method	Top-1 Accuracy			Top-5 Accuracy			Avg Class Precision			Avg Class Recall		
		Verb	Noun	Action	Verb	Noun	Action	Verb	Noun	Action	Verb	Noun	Action
Seen	Baidu-UTS 2019 [10]	69.80	52.26	41.37	90.95	76.71	63.59	63.55	46.86	25.13	46.94	49.17	26.39
	TBN Single Model [6]	64.75	46.03	34.80	90.70	71.34	56.65	55.67	43.65	22.07	45.55	42.30	21.31
	TBN Ensemble [6]	66.10	47.89	36.66	91.28	72.80	58.62	60.74	44.90	24.02	46.82	43.89	22.92
	SAP R-50(RGB)	63.22	48.34	34.76	86.10	71.48	55.91	36.98	41.94	14.60	31.56	45.24	15.94
	SAOA I3D(2-Stream)	67.58	47.79	37.68	89.21	71.83	59.25	57.79	42.13	19.62	42.65	44.75	20.72
	Ensemble w/o IG	70.13	52.49	41.78	90.97	76.71	63.92	60.20	47.38	25.00	45.40	49.57	25.84
	Ensemble w/ IG	70.41	52.85	42.57	90.78	76.62	63.55	60.44	47.11	24.94	45.82	50.02	26.93
Unseen	Baidu-UTS 2019 [10]	59.68	34.14	25.06	82.69	62.38	45.95	37.20	29.14	15.44	29.81	30.48	18.67
	TBN Single Model [6]	52.69	27.86	19.06	79.93	53.78	36.54	31.44	21.48	12.00	28.21	23.53	12.69
	TBN Ensemble [6]	54.46	30.39	20.97	81.23	55.69	39.40	32.57	21.68	10.96	27.60	25.58	13.31
	SAP R-50(RGB)	53.23	33.01	23.86	78.15	58.01	40.53	24.29	28.22	11.02	22.76	28.11	13.72
	SAOA I3D(2-Stream)	58.14	34.38	25.81	82.59	60.40	45.13	38.86	28.69	14.83	28.70	30.06	17.52
	Ensemble w/o IG	60.60	36.09	26.60	83.07	62.89	47.39	40.06	32.09	16.49	29.80	31.80	18.92
	Ensemble w/ IG	60.43	37.28	27.96	83.06	63.67	46.81	35.23	32.60	17.35	28.97	32.78	19.82

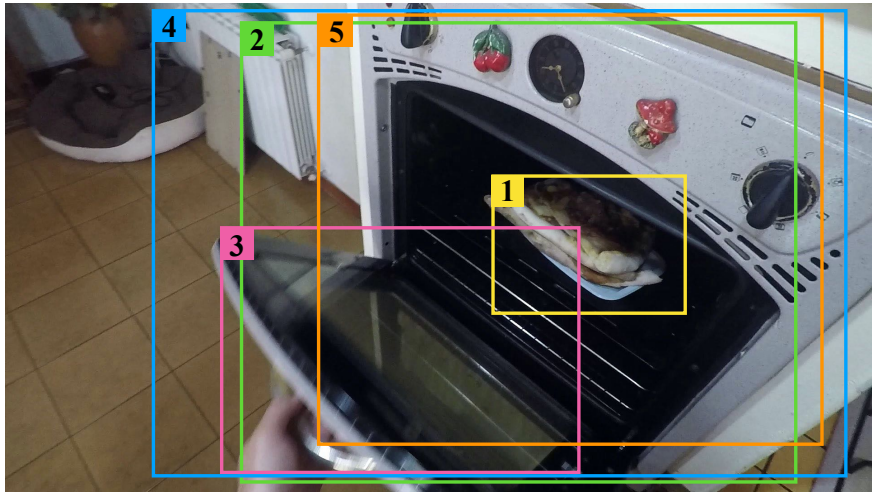
Our model achieved 1st place on both seen and unseen kitchens

Qualitative Results



	1	2	3	4	5
Noun attend Verb	8e-3	0.02	0.02	8e-3	**0.04**
Verb attend Noun	3e-3	4e-4	2e-4	2e-3	**0.78**

Verb: *'chop'* *Noun*: *'parsley'*



	1	2	3	4	5
Noun attend Verb	1e-4	0.05	**0.14**	0.06	1e-5
Verb attend Noun	3e-3	0.02	**0.16**	0.06	0.07

Verb: *'open'* *Noun*: *'oven'*

Leaderboard

EPIC-KITCHENS Action Recognition Competition 2019. 4%+ gains in top-1 accuracy compared to the team ranked 2nd.

Team Name	Top-1 Accuracy (%)			Top-5 Accuracy (%)			Precision (%)			Recall (%)		
	Verb ▲	Noun ▲	Action ▲	Verb ▲	Noun ▲	Action ▲	Verb ▲	Noun ▲	Action ▲	Verb ▲	Noun ▲	Action ▲
Baidu-UTS	69.80 (1)	52.27 (1)	41.37 (1)	90.95 (2)	76.71 (1)	63.59 (1)	63.55 (1)	46.86 (1)	25.13 (1)	46.94 (1)	49.17 (1)	26.39 (1)
Bristol-Oxford	66.10 (2)	47.88 (2)	36.66 (2)	91.28 (1)	72.80 (2)	58.62 (2)	60.73 (4)	44.89 (2)	24.01 (2)	46.81 (2)	43.88 (3)	22.92 (2)
	64.14 (4)	47.65 (3)	35.75 (3)	87.64 (5)	70.66 (4)	54.65 (6)	43.64 (14)	40.52 (5)	18.95 (9)	38.31 (9)	45.29 (2)	21.13 (5)
FBK-HUPBA	63.34 (5)	44.75 (6)	35.54 (4)	89.01 (4)	69.88 (5)	57.18 (4)	63.21 (2)	42.26 (4)	19.76 (7)	37.77 (10)	41.28 (5)	21.19 (4)
DMI-UNICT	58.99 (9)	45.00 (5)	35.14 (5)	86.70 (8)	69.08 (6)	57.62 (3)	52.23 (8)	40.06 (6)	19.40 (8)	42.12 (4)	39.32 (9)	20.28 (8)
Bristol-Oxford	64.74 (3)	46.03 (4)	34.80 (6)	90.69 (3)	71.33 (3)	56.64 (5)	55.66 (6)	43.65 (3)	22.06 (3)	45.55 (3)	42.30 (4)	21.30 (3)
NTU CML Mira	61.65 (7)	43.63 (8)	30.55 (7)	87.09 (7)	68.65 (7)	40.11 (14)	48.63 (9)	39.62 (8)	16.92 (11)	33.41 (12)	40.57 (7)	16.68 (9)

Our improved model achieved 1st in EPIC-KITCHENS Action Recognition Competition 2020.

2020 CHALLENGE WINNERS

Action Recognition

S1	S2	Team	Member	Affiliations
①	①	UTS-Baidu (wasun)	Xiaohan Wang	University of Technology Sydney, Baidu Research
			Yu Wu	University of Technology Sydney, Baidu Research
			Linchao Zhu	University of Technology Sydney
			Yi Yang	University of Technology Sydney
			Yueting Zhuang	Zhejiang University
②	③	NUS-CVML (action-banks)	Fadime Sener	University of Bonn
			Dipika Singhania	National University of Singapore
			Angela Yao	National University of Singapore
④	②	GT-WISC-MPI (aptx4869lm)	Miao Liu	Georgia Institute of Technology
			Yin Li	University of Wisconsin-Madison
			James M. Rehg	Georgia Institute of Technology
③	⑤	FBK-HUPBA (sudhakran)	Swathikiran Sudhakaran	FBK, University of Trento
			Sergio Escalera	CVC, Universitat de Barcelona
			Oswald Lanz	FBK, University of Trento
③	⑥	SAIC-Cambridge (tnet)	Juan-Manuel Perez-Rua	Samsung AI Centre, Cambridge
			Antoine Toisoul	Samsung AI Centre, Cambridge
			Brais Martinez	Samsung AI Centre, Cambridge
			Victor Escorcia	Samsung AI Centre, Cambridge
			Li Zhang	Samsung AI Centre, Cambridge
			Xiatian Zhu	Samsung AI Centre, Cambridge
			Tao Xiang	Samsung AI Centre Cambridge, Univ of Surrey

Semi-supervised video object segmentation

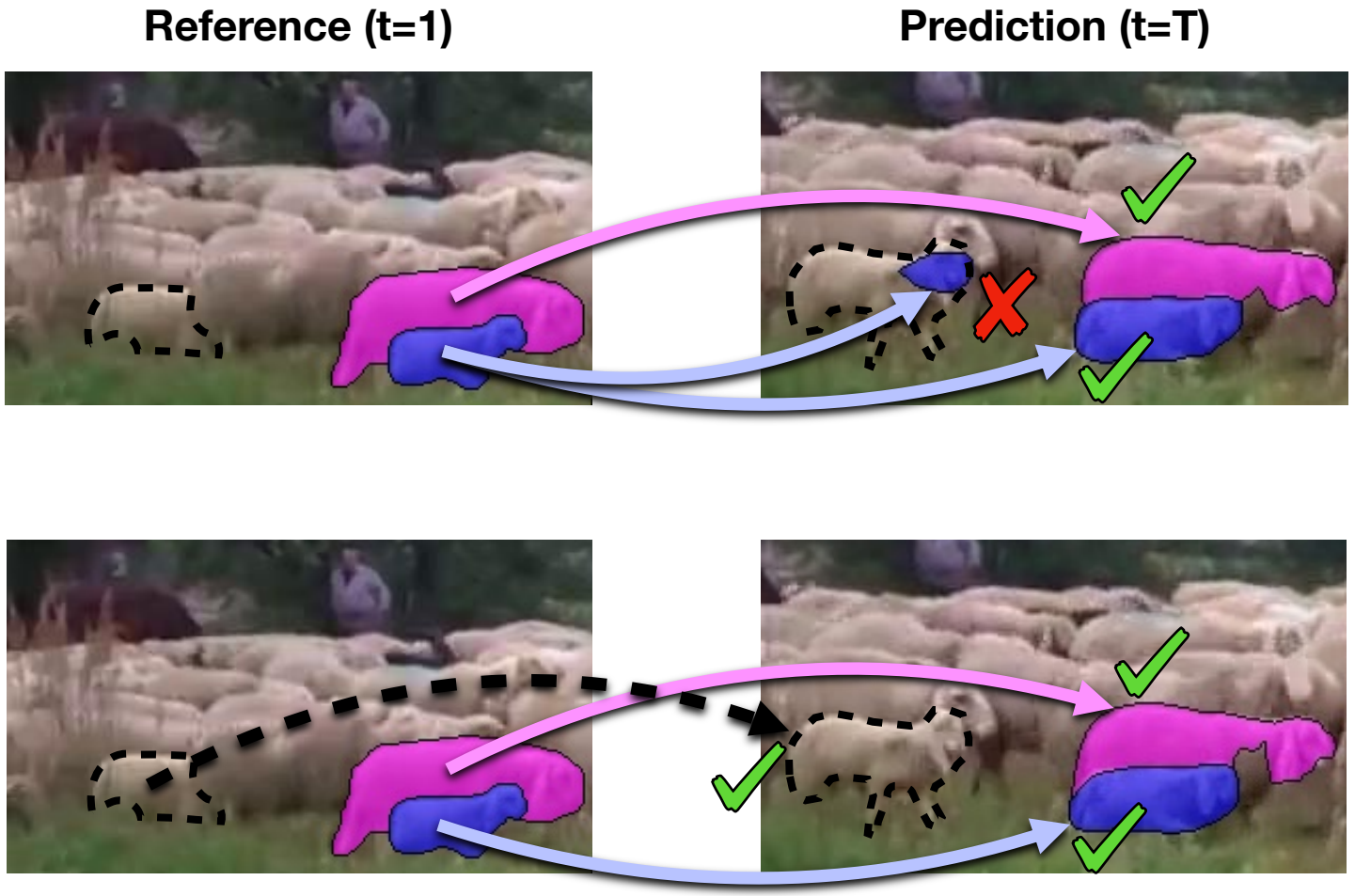
What is video object segmentation ?

- Discover and segment a primary object in a video
- With user annotations in the first frame

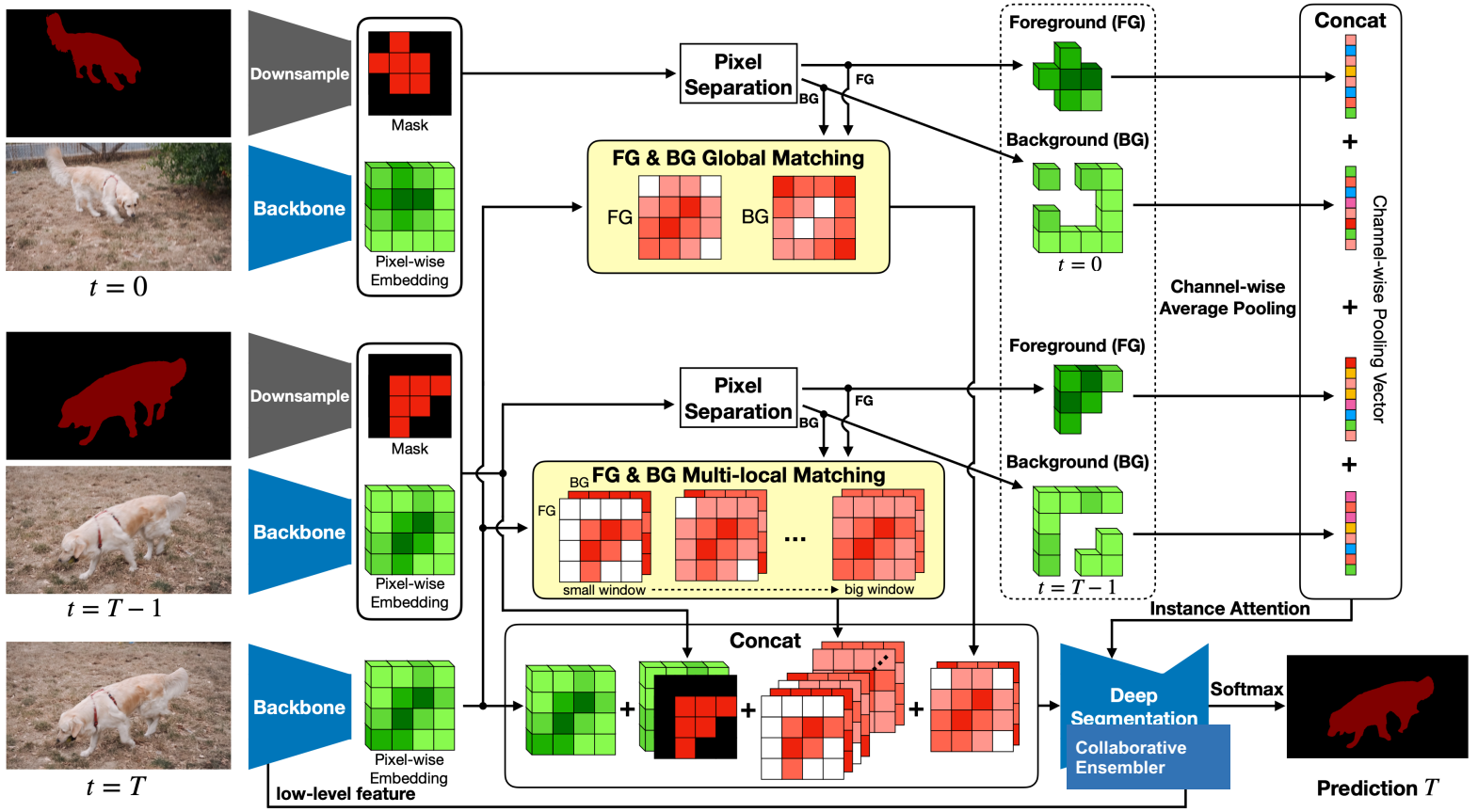
Annotated target object in the first frame

Segmentation results

Motivation: Background Matters



Collaborative VOS by Foreground-Background Integration



Collaborative Video Object Segmentation by Foreground-Background Integration. ECCV, 2020 (Spotlight)

Comparison with SOTA

Youtube-VOS

Methods	F	S	Avg	Seen $\mathcal{J}$	Seen $\mathcal{F}$	Unseen $\mathcal{J}$	Unseen $\mathcal{F}$
Validation 2018 Split							
AG [19]			66.1	67.8	-	60.8	-
PReM [22]	✓		66.9	71.4	75.9	56.5	63.7
BoLT [34]	✓		71.1	71.6	-	64.3	-
STM$^-$ [26]			68.2	-	-	-	-
STM [26]		✓	79.4	79.7	84.2	72.8	80.9
CFBI			**81.4**	**81.1**	**85.8**	**75.3**	**83.4**
CFBI$^+$			**82.7**	**82.2**	**86.8**	**76.9**	**85.0**
Testing 2019 Split							
MST* [43]		✓	81.7	80.0	83.3	**77.9**	85.5
EMN* [44]		✓	81.8	**80.7**	**84.7**	77.3	84.7
CFBI			81.5	79.6	84.0	77.3	85.3
CFBI$^+$			**82.2**	80.4	**84.7**	**77.9**	**85.7**

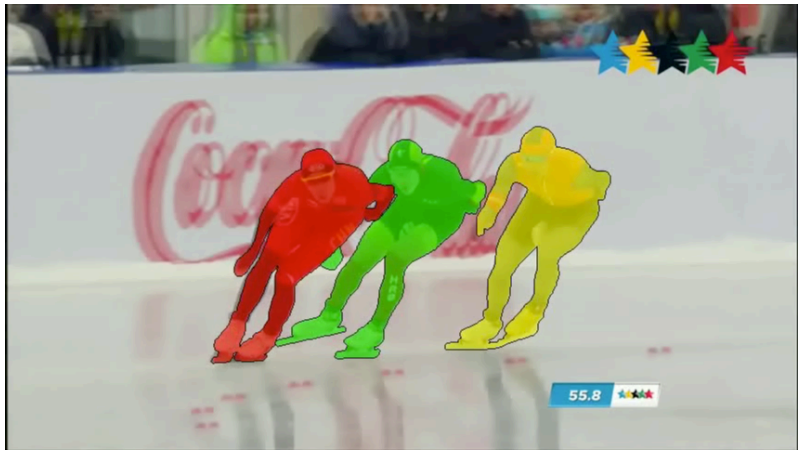
DAVIS 2016

Methods	F	S	Avg	$\mathcal{J}$	$\mathcal{F}$	t/s
OSMN [40]			-	74.0		0.14
PML [4]			77.4	75.5	79.3	0.28
VideoMatch [17]			80.9	81.0	80.8	0.32
RGMP$^-$ [36]			68.8	68.6	68.9	0.14
RGMP [36]		✓	81.8	81.5	82.0	0.14
A-GAME [19] (**Y**)			82.1	82.2	82.0	**0.07**
FEELVOS [32] (**Y**)			81.7	81.1	82.2	0.45
OnAVOS [33]	✓		85.0	85.7	84.2	13
PReMVOS [22]	✓		86.8	84.9	88.6	32.8
STMVOS [26]		✓	86.5	84.8	88.1	0.16
STMVOS [26] (**Y**)		✓	**89.3**	**88.7**	89.9	0.16
CFBI			86.1	85.3	86.9	0.18
CFBI (**Y**)			**89.4**	88.3	**90.5**	0.18
CFBI$^+$ (**Y**)			**90.7**	**89.6**	**91.7**	0

DAVIS 2017

Methods	F	S	Avg	$\mathcal{J}$	$\mathcal{F}$
Validation Split					
OSMN [40]			54.8	52.5	57.1
VideoMatch [17]			62.4	56.5	68.2
OnAVOS [33]	✓		63.6	61.0	66.1
RGMP [36]		✓	66.7	64.8	68.6
A-GAME [19] (**Y**)			70.0	67.2	72.7
FEELVOS [32] (**Y**)			71.5	69.1	74.0
PReMVOS [22]	✓		77.8	73.9	81.7
STMVOS [26]		✓	71.6	69.2	74.0
STMVOS [26] (**Y**)		✓	**81.8**	**79.2**	84.3
CFBI			74.9	72.1	77.7
CFBI (**Y**)			**81.9**	**79.1**	**84.6**
CFBI$^+$ (**Y**)			**83.3**	**80.5**	**86.0**
Testing Split					
OSMN [40]			41.3	37.7	44.9
OnAVOS [33]	✓		56.5	53.4	59.6
RGMP [36]		✓	52.9	51.3	54.4
FEELVOS [32] (**Y**)			57.8	55.2	60.5
PReMVOS [22]	✓		71.6	67.5	75.7
STMVOS [26] (**Y**)		✓	72.2	69.3	75.2
CFBI (**Y**)			**74.8**	**71.1**	**78.5**
CFBI$^+$ (**Y**)			**77.5**	**73.8**	**81.1**

Visualization



Thank you!