UCF Center for Research in Computer Vision

Stable Mean Teacher for Semi-supervised Video Action Detection

Akash Kumar, Sirshapan Mitra, Yogesh Singh Rawat

AAAI-25 / IAAI-25 / EAAI-25
FEBRUARY 25 – MARCH 4, 2025 | PHILADELPHIA, USA

SCAN ME

Why data efficient approach?

❖ **Challenges in Supervised Approach**

- ➢ Costly Training + Laborious + **Expensive**
- ➢ **Lacks** precise spatio-temporal localization
- ➢ **No** temporal coherency

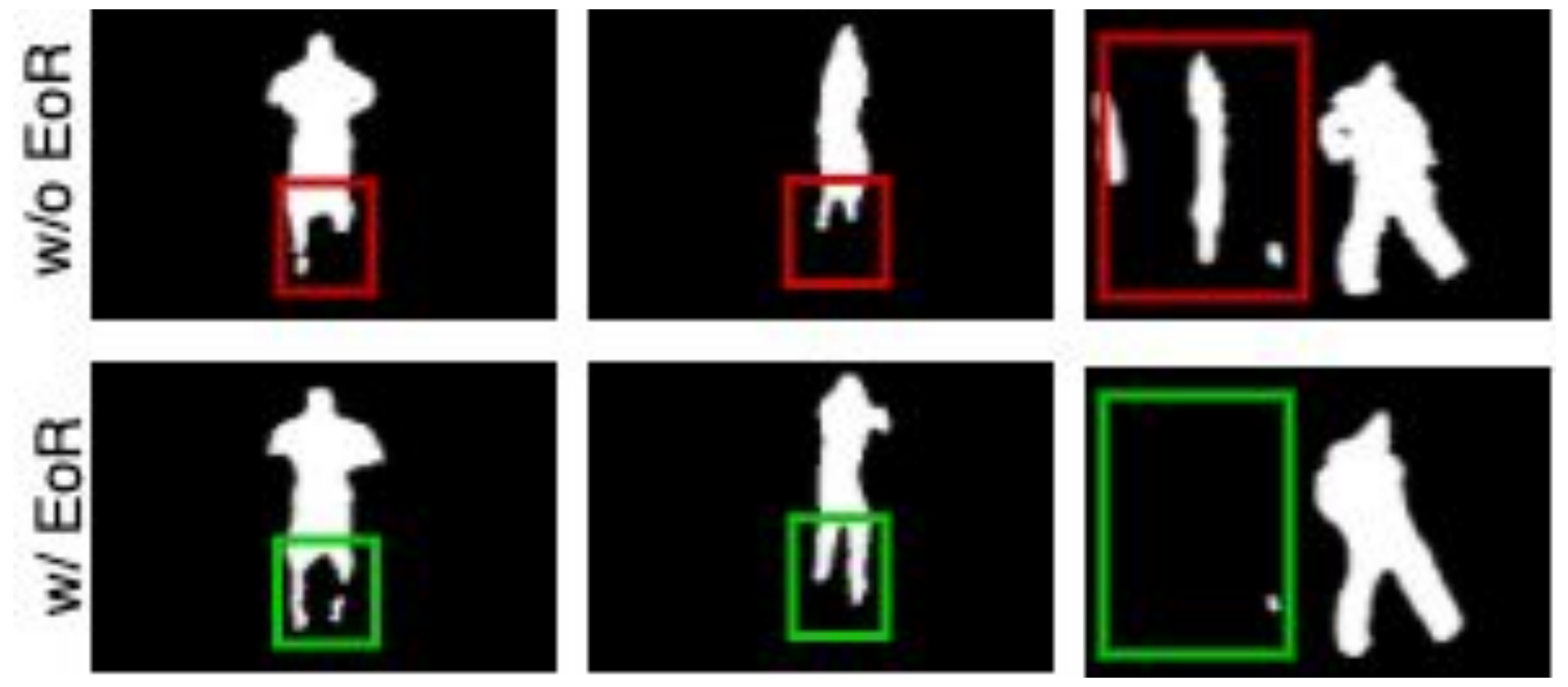
❖ **Solutions**

- ➢ Label **Efficient** Approach
- ➢ **Recover** fine-grained level localization mistake
- ➢ Enforce temporal **smooth** flow

Contributions

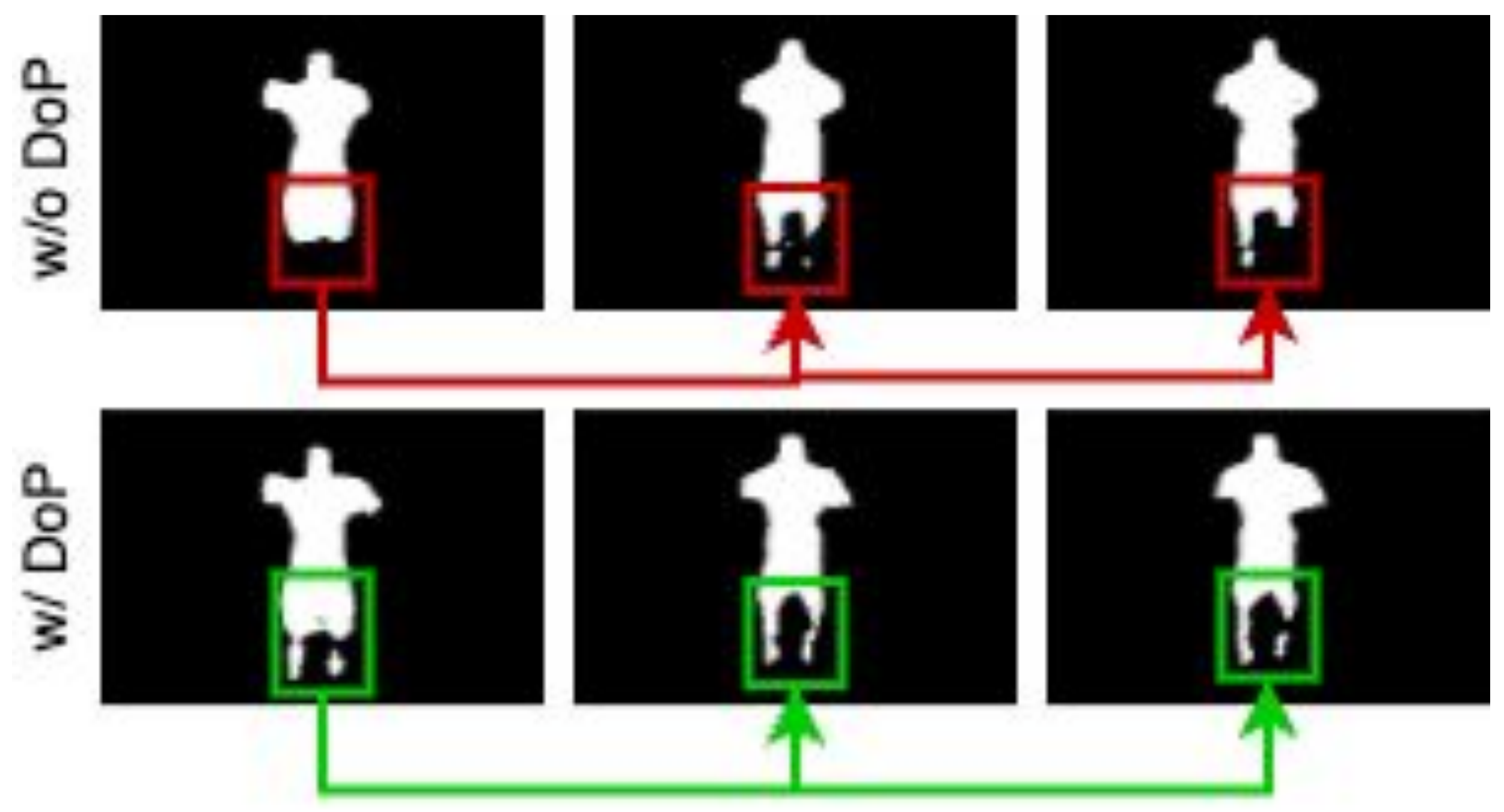
❖ **Error Recovery (EoR)**

- ➢ **Aim:** Spatial boundary refinement
- ➢ How?
  - ■ Learn from **student's mistake**
  - ■ Provide teacher with better supervisory signal

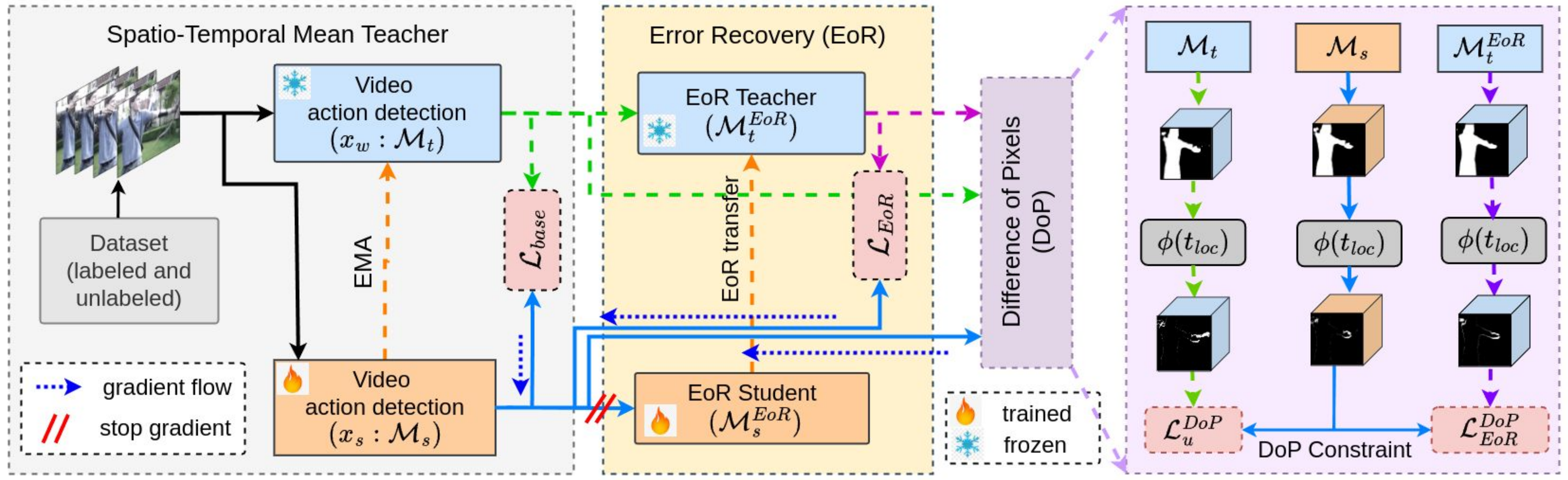


❖ **Difference of Pixels (DoP)**

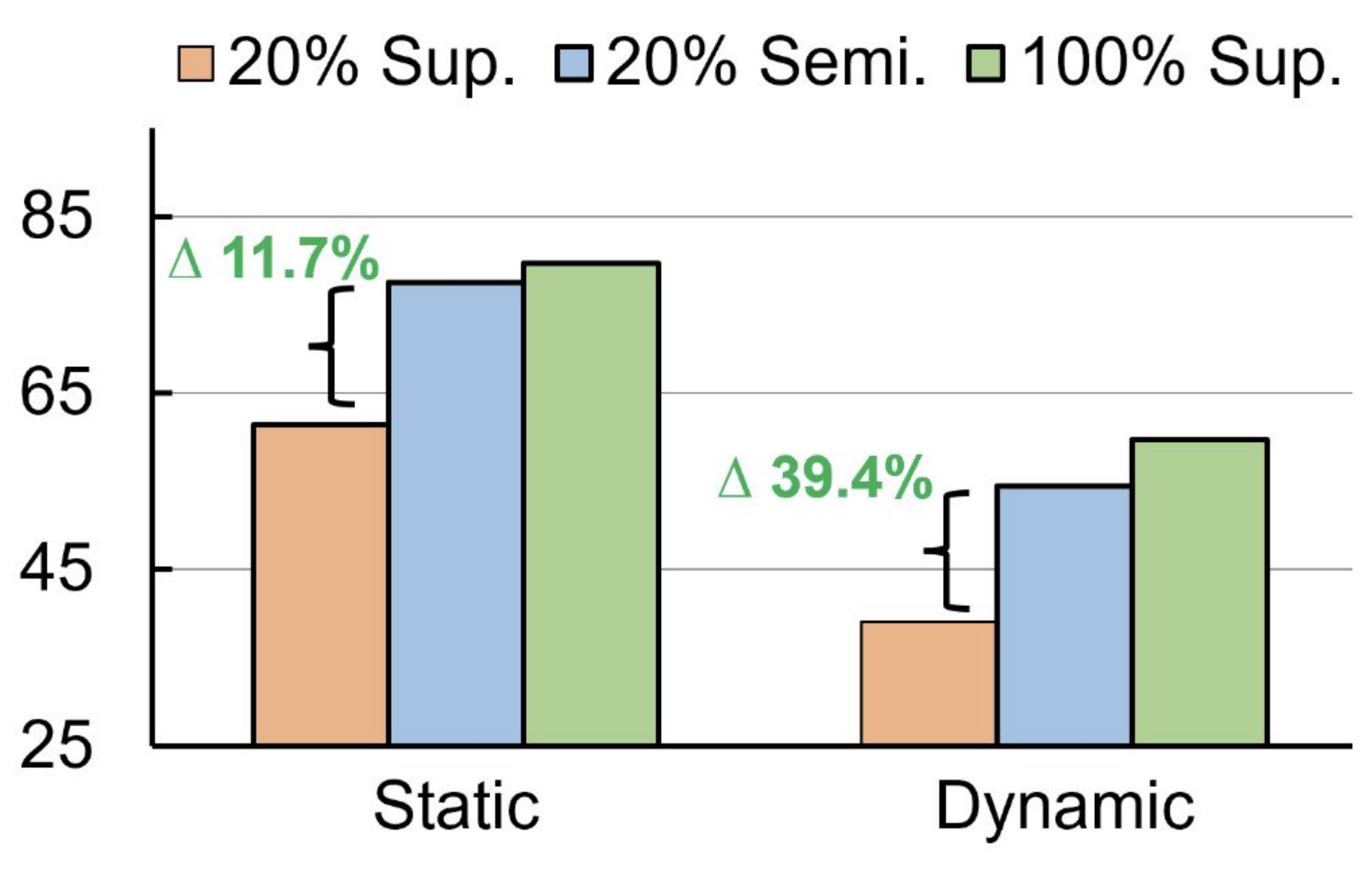
- ➢ **Aim:** Spatio-Temporal coherency induction
- ➢ How?
  - ■ Optimization of **pixel difference** across time

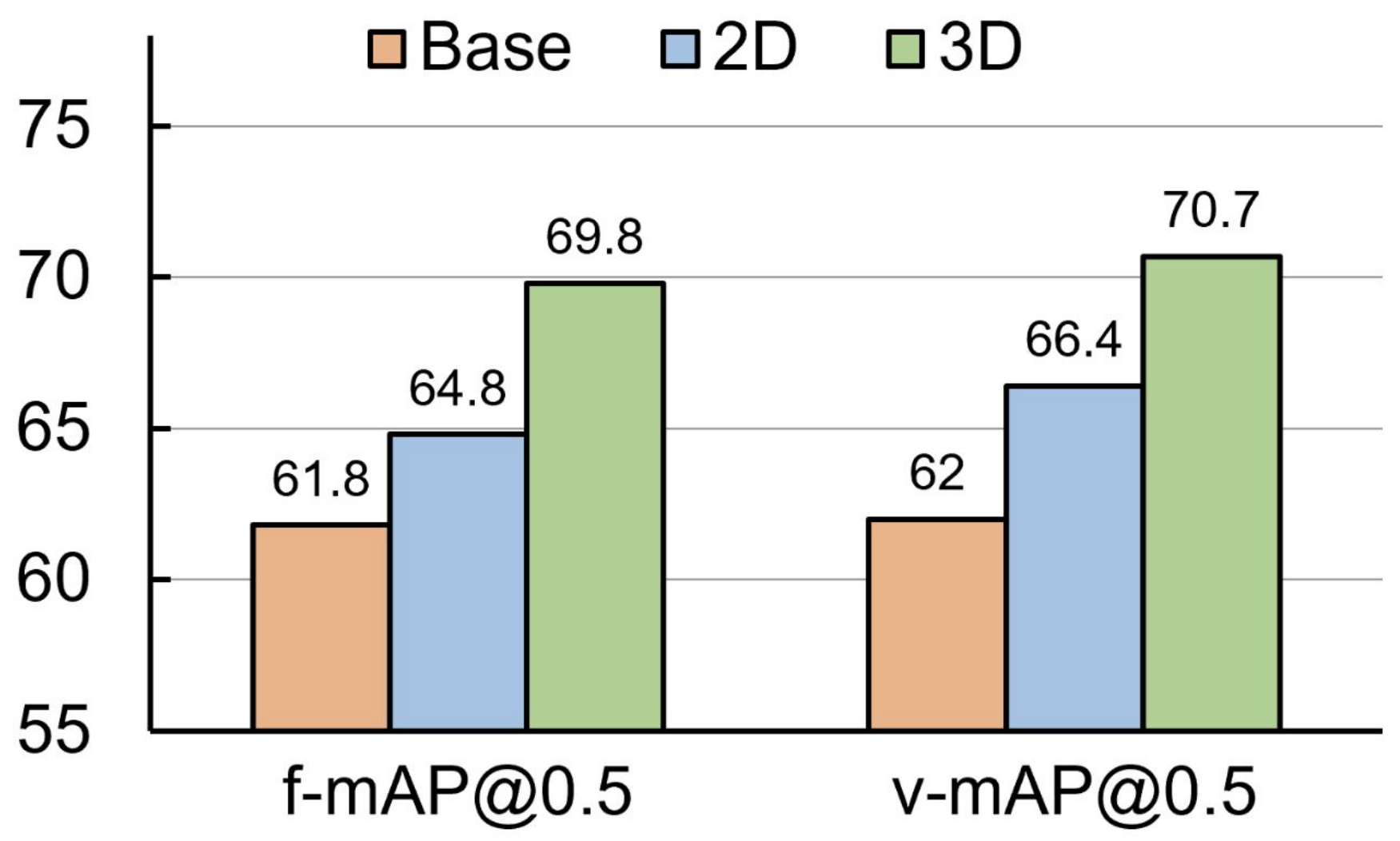


Proposed Approach


Spatio-Temporal Mean Teacher
Dataset (labeled and unlabeled)
Video action detection $(x_w : \mathcal{M}_t)$
EMA
Video action detection $(x_s : \mathcal{M}_s)$
$\mathcal{L}_{base}$
gradient flow
stop gradient
Error Recovery (EoR)
EoR Teacher $(\mathcal{M}_t^{EoR})$
EoR transfer
$\mathcal{L}_{EoR}$
EoR Student $(\mathcal{M}_s^{EoR})$
Difference of Pixels (DoP)
trained
frozen
$\mathcal{M}_t$
$\mathcal{M}_s$
$\mathcal{M}_t^{EoR}$
$\phi(t_{loc})$
$\phi(t_{loc})$
$\phi(t_{loc})$
$\mathcal{L}_u^{DoP}$
DoP Constraint
$\mathcal{L}_{EoR}^{DoP}$

Analysis





★ Dynamic → ↑ challenging

★ Dynamic > Static (+ Δ 27 %)

★ Error Recovery Architecture
3D > 2D

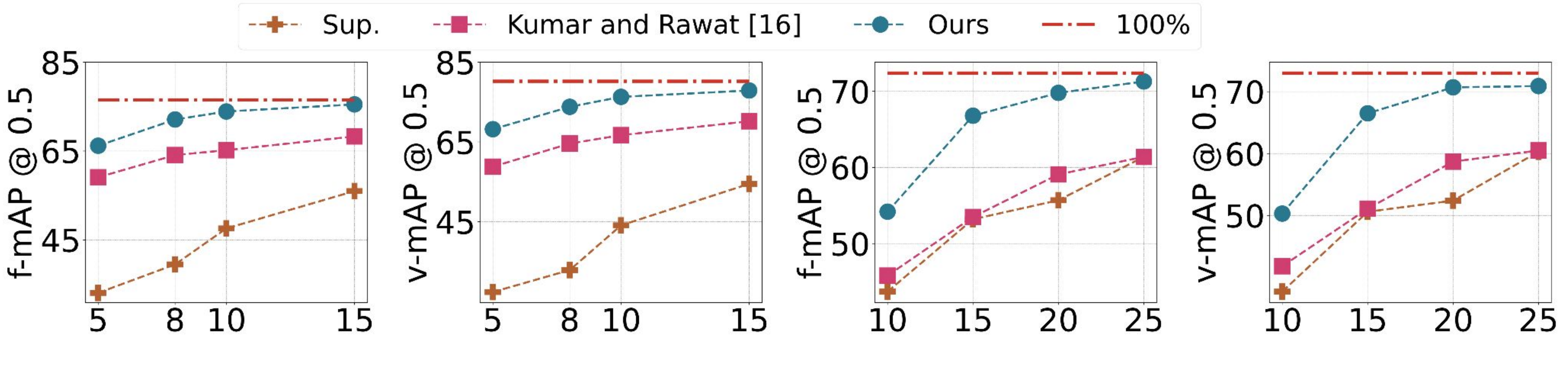
Results

Semi-Supervised Approaches	Backbone	UCF101-24 Annot.	UCF101-24 f@0.5	UCF101-24 v@0.2	UCF101-24 v@0.5	JHMDB21 Annot.	JHMDB21 f@0.5	JHMDB21 v@0.2	JHMDB21 v@0.5
MixMatch (Berthelot et al. 2019)††	I3D	10%	10.3	54.7	4.9	30%	7.5	46.2	5.8
Pseudo-label (Lee et al. 2013)	I3D	10%	59.3	89.9	58.3	20%	55.3	87.6	52.0
ISD (Jeong et al. 2021)	I3D	10%	60.2	91.3	64.0	20%	57.8	90.2	57.0
E2E-SSL (Kumar and Rawat 2022)	I3D	10%	65.2	91.8	66.7	20%	59.1	93.2	58.7
Mean Teacher (Tarvainen and Valpola 2017)	I3D	10%	67.3	92.7	70.5	20%	56.3	88.8	52.8
Stable Mean Teacher (Ours)	I3D	10%	**73.9**	**95.8**	**76.3**	20%	**69.8**	**98.8**	**70.7**
			(↑ 6.6)	(↑ 3.1)	(↑ 5.8)		(↑ 13.5)	(↑ 10.0)	(↑ 17.9)

Scaling to large-scale dataset (AVA)

Method	Backbone	Pretraining	F	FPS	$\mathcal{A}$	mAP	GFLOPs
Real-time spatio-temporal action detector							
YOWO (2019)	ResNext-101	K400	16	35	100%	17.9	44
YOWOv2-N (2023)	Shufflev2-1.0x	K400	16	40	100%	12.6	1.3
Ours(YOWOv2-N)	Shufflev2-1.0x	K400	16	40	10%	8.5	1.3
Sup. baseline	Shufflev2-1.0x	K400	16	40	10%	5.2	1.3

Comparison at different annotation percentages



UCF101-24

JHMDB-21

Generalization (Video Object Segmentation)

Method	Annot.	Avg	J_S	J_U	F_S	F_U
Xu (2018b)	100%	47.9	55.7	39.6	55.2	41.3
Xu (2018b) †	10%	10.1	11.6	10.1	9.6	9.2
Kumar *et al.* (2022)	10%	36.8	43.1	31.4	40.8	31.8
Ours	10%	41.3	48.2	35.0	46.7	35.4
		(↑ 4.5)	(↑ 5.1)	(↑ 3.6)	(↑ 5.9)	(↑ 3.6)

Qualitative Analysis