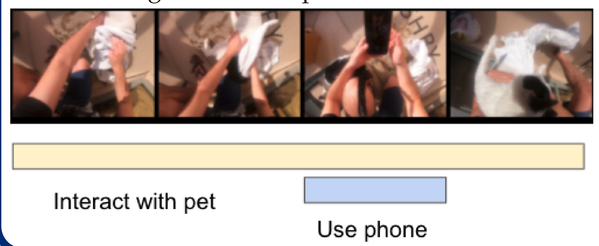
Ego-Only: Egocentric Action Detection without Exocentric Transferring

Huiyu Wang[1] Mitesh Kumar Singh[1] Lorenzo Torresani[1]
[1]Meta AI

Meta AI

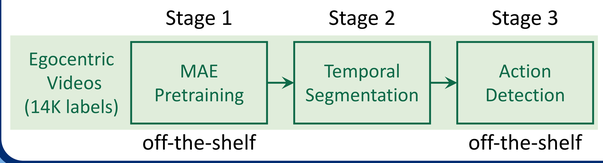
Goal

- Egocentric action detection and recognition.
  a.k.a. egocentric temporal action localization.



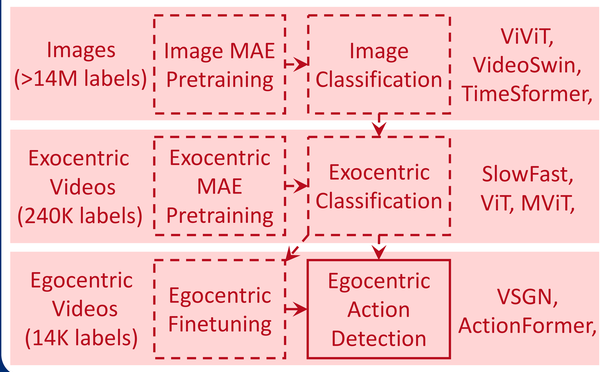
Method

- **Our Ego-Only approach** learns from ego videos only and simplifies the existing pipeline that uses extra data, extra labels, and extra pretraining stages.



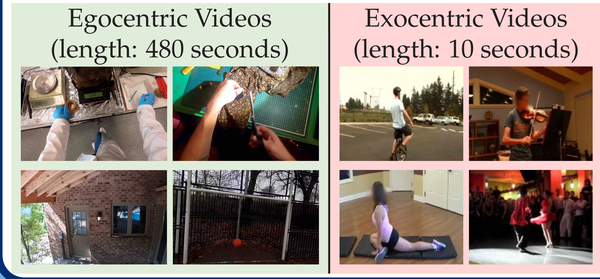
Previous Methods

- **Previous methods** depend heavily on transferring from exo to ego, despite the content and appearance gap separating the two domains.



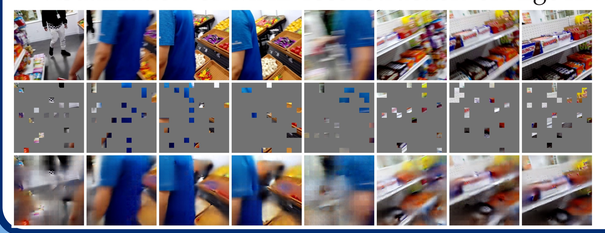
Ego-Exo Domain Gap

- Egocentric domain specific challenges:
  - No actor visible but only hand object interaction.
  - Fine-grained actions, long-tail class distribution.
  - Domain shift between ego and exo datasets.
  - Localization in long-form videos.

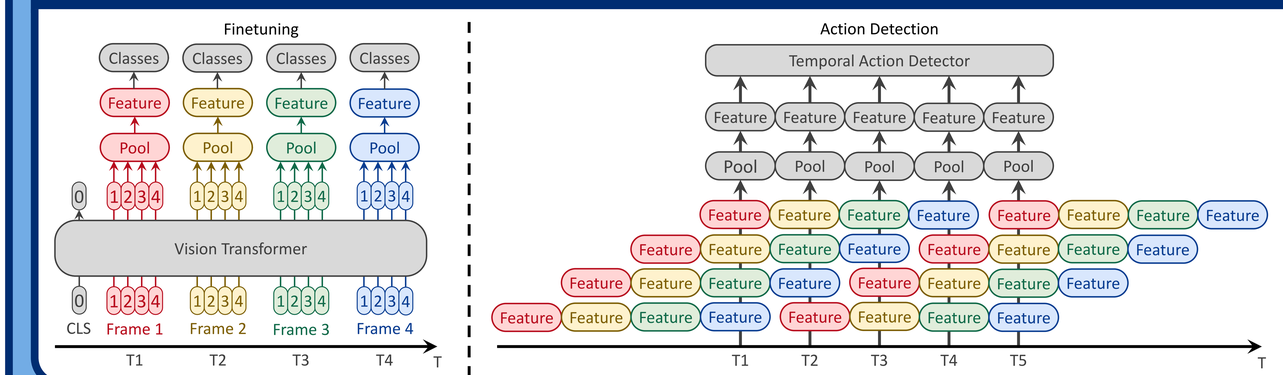


Stage 1 MAE Pretraining

- Visualization of MAE reconstruction on Ego4D.

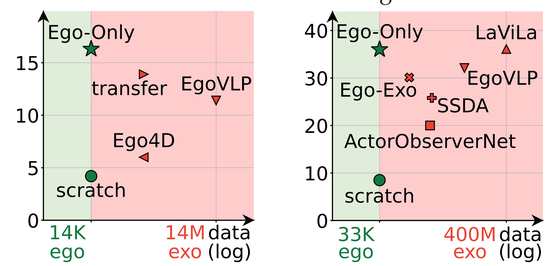


Stage 2 Temporal Segmentation & Stage 3 Action Detection



Main Results

- Ego-Only uses **orders of magnitude fewer labels**, simplifies the pipeline, and improves the results.



- EPIC-Kitchens-100 Action Detection mAP.

method	backbone	exo data	verb	noun	# labels
ActionFormer	SlowFast	K400	23.5	21.9	307K
Ego-Only	ViT-B	-	**28.1**	**26.5**	**67K**

- EPIC-Kitchens-100 Recognition top-1 accuracy.

method	variant	verb	noun
ViViT	ViViT-L/16x2, IN-21k+K400	66.4	56.8
MoViNet	MoViNet-A6, 120 frames	72.2	57.3
MeMViT	MeMViT, 32×3, K600, 105.6 sec	71.4	60.3
LaViLa	TSF-L, WebImageText+Ego4D	72.0	62.9
Ego-Only	ViT-L, 32 frames, 3.2 sec	**73.3**	59.4

Ablation Study

- Varying the pretraining stage with Video MAE.

method	self-sup. MAE	sup. exo	sup. ego	Ego4D mAP	# labels seen
exo-sup	-	K400	Ego4D	13.9	254K (18×)
ours	Ego4D	-	Ego4D	**16.3**	**14K (1×)**
scratch	-	-	Ego4D	4.2	14K (1×)
exo-MAE	K400	-	Ego4D	13.4	14K (1×)
exo-FT	K400	K400	Ego4D	16.2	254K (18×)

- Varying the finetuning stage with segmentation.

method	backbone	self-sup. MAE	sup. exo	sup. ego	Ego4D mAP
exo-sup	SlowFast	-	K400	-	13.2
exo-MAE	ViT-B	K400	-	-	6.7
ego-MAE	ViT-B	Ego4D	-	-	7.8
exo-FT	ViT-B	K400	K400	-	13.5
ours	ViT-B	Ego4D	-	Ego4D	16.3

- Joint ego-exo pretraining on Ego4D and Kinetics.

method	self-sup. MAE	sup. exo	sup. ego	Ego4D mAP
exo-MAE	K400	-	Ego4D	13.4
joint-MAE	K400 & Ego4D	-	Ego4D	16.0
ours	Ego4D	-	Ego4D	**16.3**

- Joint ego-exo finetuning on KEEC.
  - **K**inetics-600, **E**go4D, **E**PIC-Kitchens-100, **C**OIN.

method	self-sup. MAE	sup. exo	sup. ego	verb mAP	noun mAP	# labels seen
joint-FT	EPIC	-	KEEC	28.4	27.9	515K
ours	EPIC	-	EPIC	**29.0**	**28.1**	**67K**

Runtime

method	FLOPs (G) MAE	exo	ego	training time (hours) MAE	exo	ego	total
exo-sup	-	598	598	-	200.9	10.0	210.9
exo-FT	81	598	598	100.5	50.2	10.0	160.7
ours	81	-	598	100.5	-	10.0	**110.5**