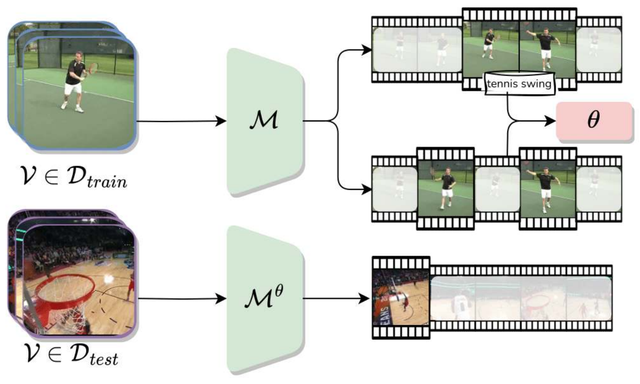
Test-Time Zero-Shot Temporal Action Localization

Benedetta Liberatori[1] Alessandro Conti[1] Paolo Rota[1] Yiming Wang[2] Elisa Ricci[1,2]

University of Trento [1] Fondazione Bruno Kessler [2]

Motivation

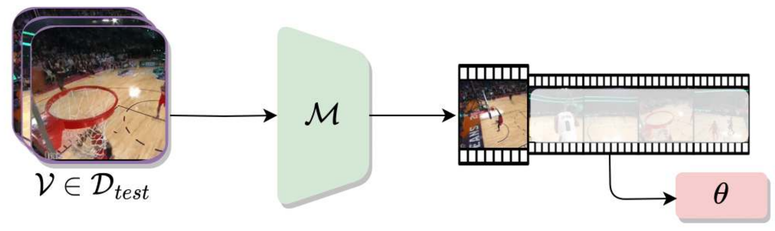
Previous works require training data to learn to localize actions



Main Problems

assume the availability of a large annotated data collection

poorly generalise out-of-distribution

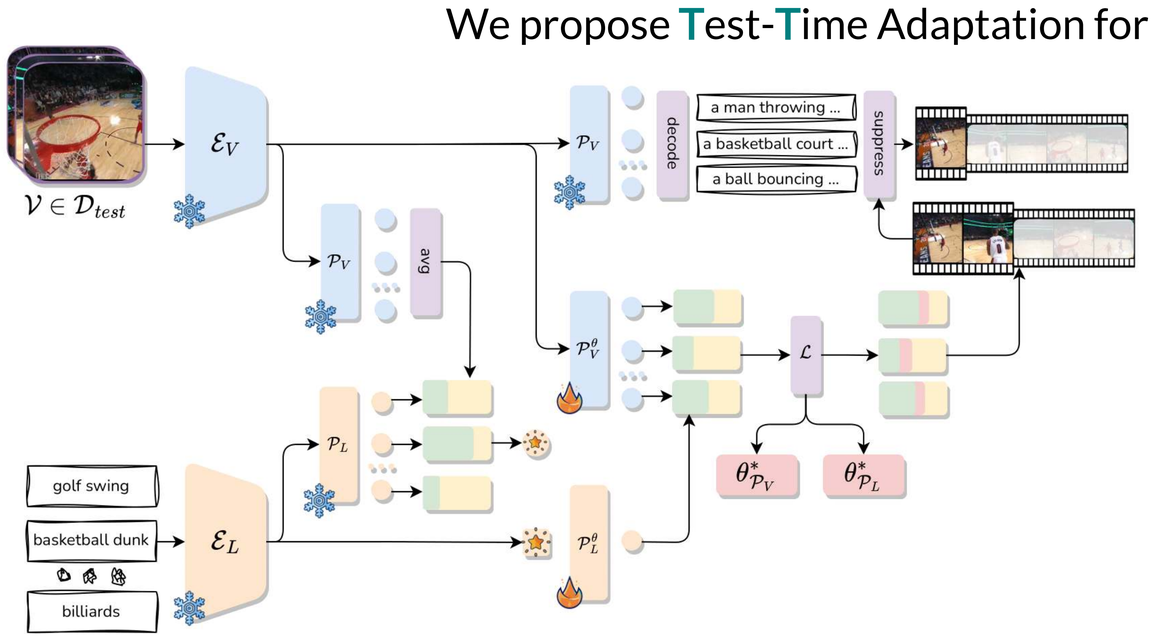


Our method updates the parameters at test-time on a stream of unlabelled videos without prior supervised training

Can we classify & localize actions in untrimmed videos without training data?

Method

We propose **T**est-**T**ime Adaptation for **T**emporal **A**ction **L**ocalization (**T3AL**)



Cosine similarity between the textual and the averaged visual embeddings to obtain a video pseudo-label

Refine the (visual embeddings -) similarities with TTA

Fine-tune the vision and language projectors, keeping the encoders frozen

Generate captions and perform text-guided region suppression

Predict & reinitialize the parameters

Results

Method	mAP (%) ↑ 0.50	0.75	0.95	Avg.
CLIP$_{32}$ [22]	0.4	0.2	0.0	0.2
CLIP$_{16}$ [22]	0.8	0.3	0.0	0.3
CoCa [31]	2.3	1.0	0.2	1.1
$T3AL_{T=0}$	24.2	13.0	2.8	13.3
$T3AL$	25.8	13.9	3.1	14.3
CLIP$_{16}$ w/ Detector [9, 20]	28.0	16.4	1.2	16.0
EffPrompt [9]	32.0	19.3	2.9	19.6
STALE [20]	32.1	20.7	5.9	20.5
UnLoc [30]	43.7	-	-	-

Table 3. **Results on ActivityNet-v1.3 (50%-50%).** Green is our method, purple indicates training-based approaches.

Method	mAP (%) ↑ 0.50	0.75	0.95	Avg.
CLIP$_{32}$ [22]	0.4	0.1	0.0	0.2
CLIP$_{16}$ [22]	0.9	0.3	0.1	0.4
CoCa [31]	3.1	1.3	0.3	1.6
$T3AL_{T=0}$	26.1	13.9	2.9	14.3
$T3AL$	28.1	14.9	3.3	15.4
CLIP$_{16}$ w/ Detector [9, 20]	35.6	20.4	2.1	20.2
EffPrompt [9]	37.6	22.9	3.8	23.1
STALE [20]	38.2	25.2	6.0	24.9
UnLoc [30]	48.8	-	-	-

Table 4. **Results on ActivityNet-v1.3 (75%-25%).** Green is our method, purple indicates training-based approaches.

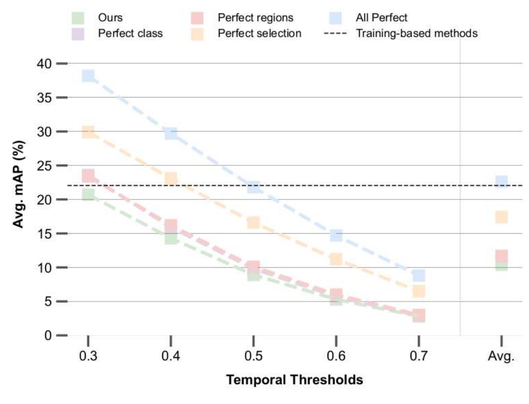
Method	mAP (%) ↑ 0.3	0.4	0.5	0.6	0.7	Avg.
CLIP$_{32}$ [22]	7.2	4.1	2.3	1.1	0.5	3.0
CLIP$_{16}$ [22]	7.5	4.2	2.2	1.1	0.6	3.1
CoCa [31]	8.4	4.7	2.5	1.2	0.6	3.5
$T3AL_{T=0}$	11.4	6.8	3.5	1.7	0.6	4.8
$T3AL$	20.7	14.3	8.9	5.3	2.7	10.4
CLIP$_{16}$ w/ Detector [9, 20]	27.2	21.3	15.3	9.7	4.8	15.7
EffPrompt [9]	37.2	29.6	21.6	14.0	7.2	21.9
STALE [20]	38.3	30.7	21.2	13.8	7.0	22.2

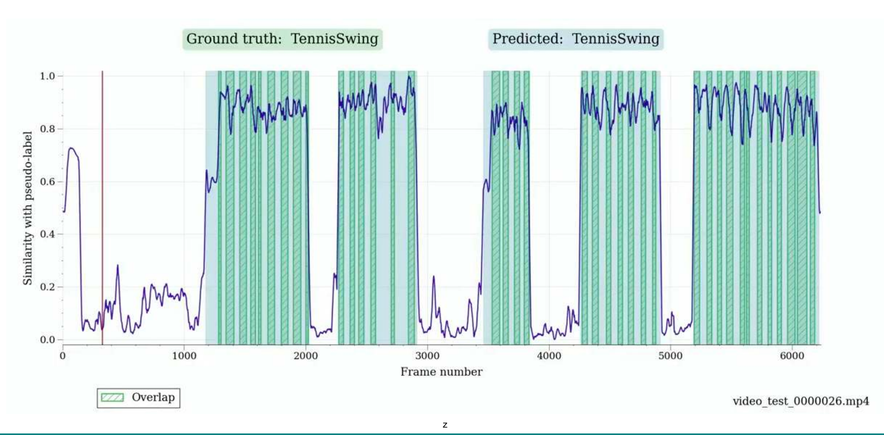
Table 1. **Results on THUMOS14 (50%-50%).** Green is our method, purple indicates training-based approaches.

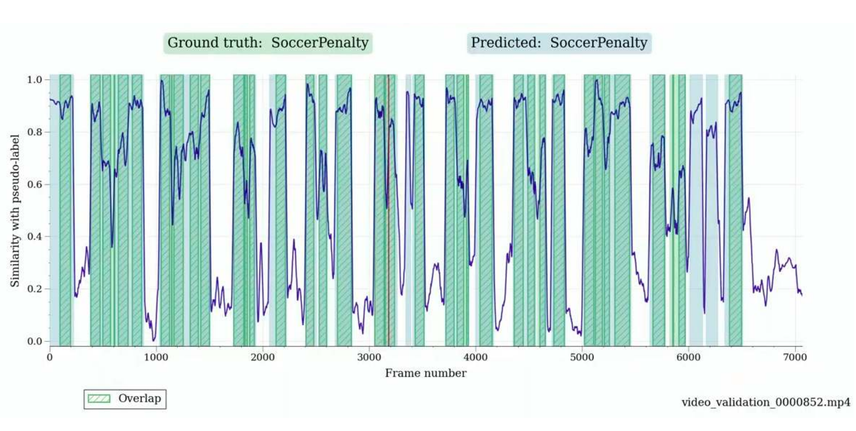
Method	mAP (%) ↑ 0.3	0.4	0.5	0.6	0.7	Avg.
CLIP$_{32}$ [22]	5.5	3.3	1.9	0.9	0.4	2.4
CLIP$_{16}$ [22]	6.9	3.8	2.1	1.1	0.6	2.9
CoCa [31]	7.8	4.6	2.5	1.3	0.6	3.4
$T3AL_{T=0}$	11.1	6.5	3.2	1.5	0.6	4.6
$T3AL$	19.2	12.7	7.4	4.4	2.2	9.2
CLIP$_{16}$ w/ Detector [9, 20]	33.0	25.5	18.3	11.6	5.7	18.8
EffPrompt [9]	39.7	31.6	23.0	14.9	7.5	23.3
STALE [20]	40.5	32.3	23.5	15.3	7.6	23.8

Table 2. **Results on THUMOS14 (75%-25%).** Green is our method, purple indicates training-based approaches.

We re-evaluate T3AL with partial **oracle** information as perfect class prediction for the pseudo-label, perfect regions count selection and perfect selection







Contributions

Zero-shot temporal action localization (ZS-TAL) in a **new practical scenario** where training data is unavailable

T3AL: a ZS-TAL method that benefits from an effective test-time adaptation (TTA) strategy and requires no training data

Adapting on an unlabeled stream of data is a **viable solution to the out-of-distribution issue** of current training-based approaches for ZS-TAL