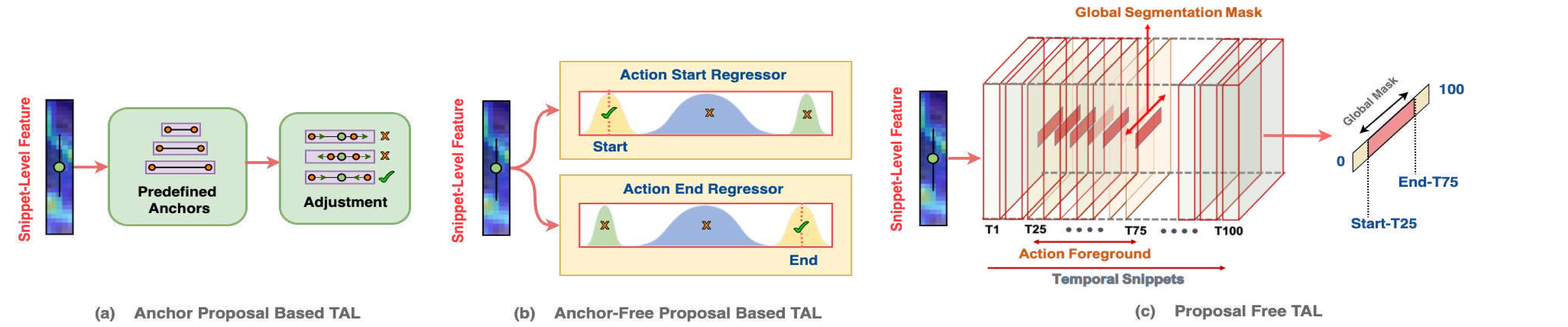
Proposal-Free Temporal Action Detection via Global Segmentation Mask Learning

Sauradip Nag, Xiatian Zhu, Yi-Zhe Song, Tao Xiang
s.nag@surrey.ac.uk https://github.com/TAGS
UNIVERSITY OF SURREY
People-Centred AI UNIVERSITY OF SURREY
iFLYTEK

ECCV TEL AVIV 2022

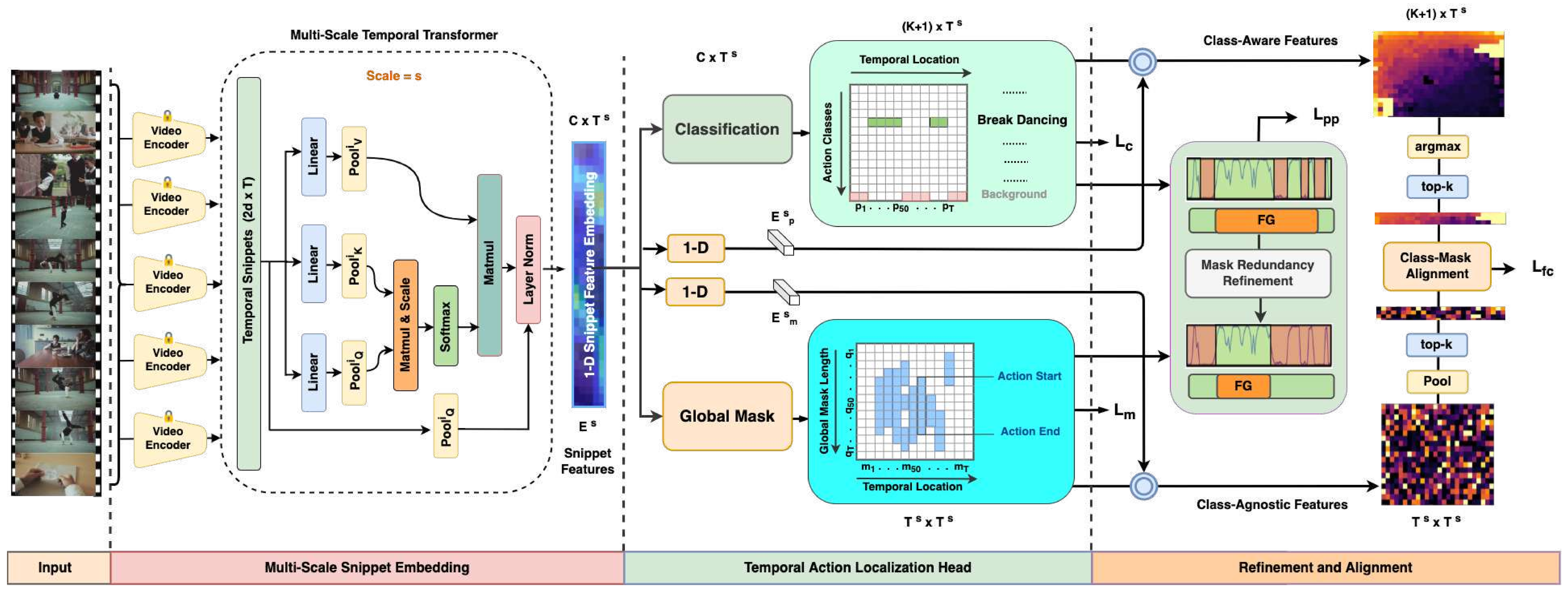
Introduction

Background: Temporal action detection (TAD) aims to identify the temporal interval (i.e., the start and end points) and the class label of all action instances in an untrimmed video.
Motivation: All existing TAD methods rely on proposal generation by either regressing predefined anchor boxes (Fig. 1(a)) or directly predicting the start and end times of proposals (Fig. 1(b)). It takes a local view of the video and focus on each individual proposal for refinement and classification. It has some limitations: (1) An excessive (sometimes exhaustive) number of proposals are usually required for good performance. - *cost ineffective* (2) Once the proposals are generated, the subsequent modeling is local to each individual proposal - *missing global context*.
Contributions: (1) Proposed a novel *proposal-free TAD model* based on global segmentation mask (TAGS) learning with simpler design and low computation cost; (2) To enhance the learning of temporal boundary, we proposed a novel boundary focused loss, along with mask predictive redundancy; (3) SOTA performance on ActivityNet and THUMOS.



(a) Anchor Proposal Based TAL (b) Anchor-Free Proposal Based TAL (c) Proposal Free TAL

Model Architecture



Learning Objectives

Softmax Cross Entropy (CE)

classifying the temporally dependent snippet specific action classes

$$\mathcal{L}_{\mathrm{SC}} = \lambda_1 (1 - p(y))^{\gamma} \log(p_y)$$

Classification Regression (CR)

classifying the snippets independently using sigmoid which also models the masks in class branch

$$\mathcal{L}_{\mathrm{CR}} = (1-\lambda_1)\big(\log(r_y) - \frac{\alpha}{|\mathcal{N}|}\sum_{k \epsilon \mathcal{N}} (\log(1 - r(k)))\big)$$

Boundary IOU (bIOU)

calculates IOU of action mask boundaries and also penalizes for no overlap of boundaries

$$\mathcal{L}_{\mathrm{bIOU}} = 1 - \Big(\frac{\cap(m,g)}{\cup(m,g)} + \frac{1}{\cap(m,g)+\epsilon}\frac{\|m-g\|_2}{c}\Big)$$

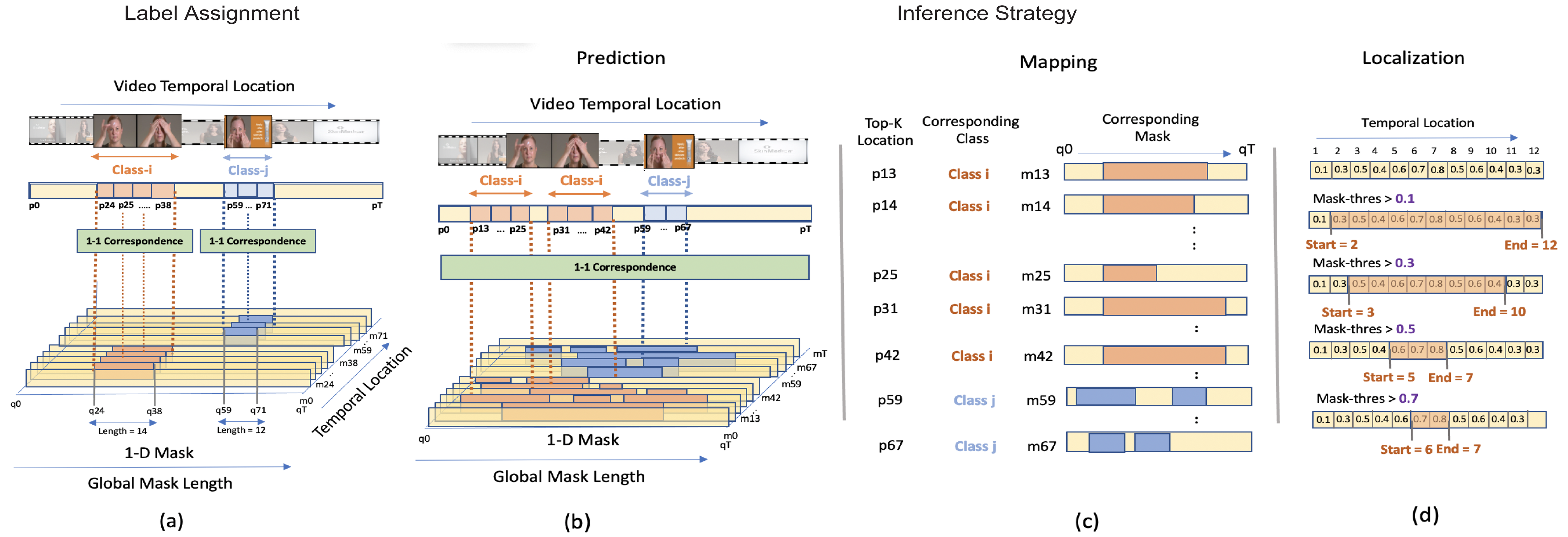
Mask Redundancy (MR)

estimates the inter-branch prediction redundancy and mask branch per-instance mask consistency

$$\mathcal{L}_{\mathrm{MR}} = \big(1 - \mathbb{R}(\pi[j^*])\big)^{\beta} \|m_t - g_t\|_2 + \mathcal{L}_{\cos},$$

Label Assignment and Inference

GT Label Assignment : To train TAGS, the ground-truth needs to be arranged into the designed format. (1) We label all the snippets (orange or blue squares) of a single action instance with the same action class. (2) For an action snippet, its global mask is defined as the video-length binary mask of that action instance. (3) Each mask is action instance specific and all snippets of a action instance share the same mask.
Inference: Given a test video, we start with the top $\%M_1$ scoring snippets from class branch (Fig 3(b)), we obtain their segmentation mask predictions (Fig 3(c)) by thresholding the corresponding columns of mask branch (Fig 3(d)). We then combine the scores and use SoftNMS for post-processing.



(a) (b) (c) (d)

Main Results

Results on ActivityNetv1.3 and THUMOS14

Type	Model	Bkb	THUMOS14 0.3	0.4	0.5	0.6	0.7	Avg.	ActivityNet-v1.3 0.5	0.75	0.95	Avg.
Anchor	R-C3D	C3D	44.8	35.6	28.9	-	-	-	26.8	-	-	-
	GTAN	P3D	57.8	47.2	38.8	-	-	-	52.6	34.1	8.9	34.3
	MUSES	I3D	**68.9**	**64.0**	56.9	46.3	31.0	**53.4**	50.0	34.9	6.5	34.0
Anchor-free	BMN	TS	56.0	47.4	38.8	29.7	20.5	38.5	50.1	34.8	8.3	33.9
	G-TAD	TS	54.5	47.6	40.2	30.8	23.4	39.3	50.4	34.6	9.0	34.1
	BU-TAL	I3D	53.9	50.7	45.4	38.0	28.5	43.3	43.5	33.9	9.2	30.1
	TCANet	TS	60.6	53.2	44.6	36.8	26.7	-	52.2	36.7	6.8	35.5
	ContextLoc	I3D	68.3	63.8	54.3	41.8	26.2	-	56.0	35.2	3.5	34.2
	RTD-Net	I3D	68.3	62.3	51.9	38.8	23.7	-	47.2	30.7	8.6	30.8
Proposal-Free	TAGS (Ours)	I3D	68.6	63.8	**57.0**	**46.3**	**31.8**	52.8	**56.3**	**36.8**	**9.6**	**36.5**
	TAGS (Ours)	TS	**61.4**	**52.9**	46.5	38.1	27.0	**44.0**	**53.7**	36.1	**9.5**	**35.9**

Ablation Studies

Analysis of model training and test cost.

Model	Epoch	Train	Test
BMN	13	6.45 hr	0.21 sec
G-TAD	11	4.91 hr	0.19 sec
TAGS	**9**	**0.26** hr	**0.12** sec

Analysis of model parameters # and FLOPs.

Model	Params (in M)	FLOPs (in G)
BMN	5.0	91.2
GTAD	9.5	97.2
TAGS	**6.2**	**17.8**

False Positive Analysis.