Semi-Supervised Temporal Action Detection with Proposal-Free Masking

Sauradip Nag, Xiatian Zhu, Yi-Zhe Song, Tao Xiang

s.nag@surrey.ac.uk https://github.com/SPOT

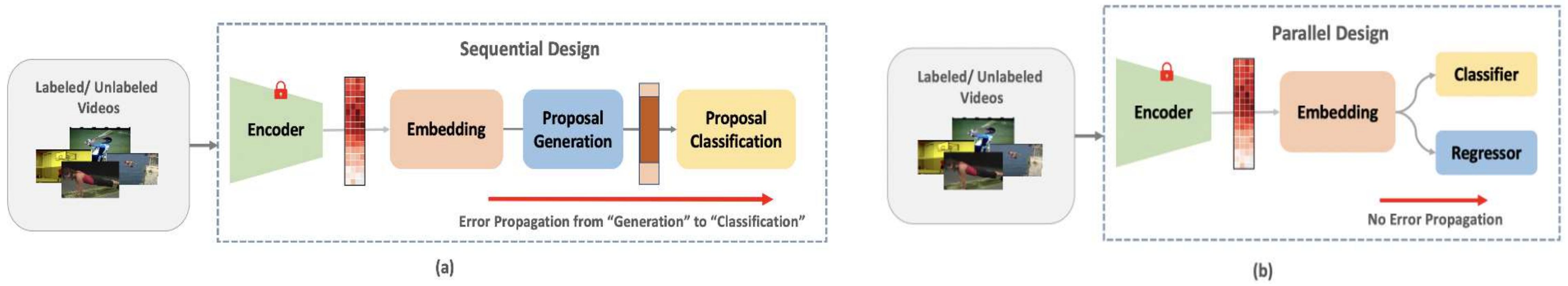
ECCV TEL AVIV 2022

Introduction

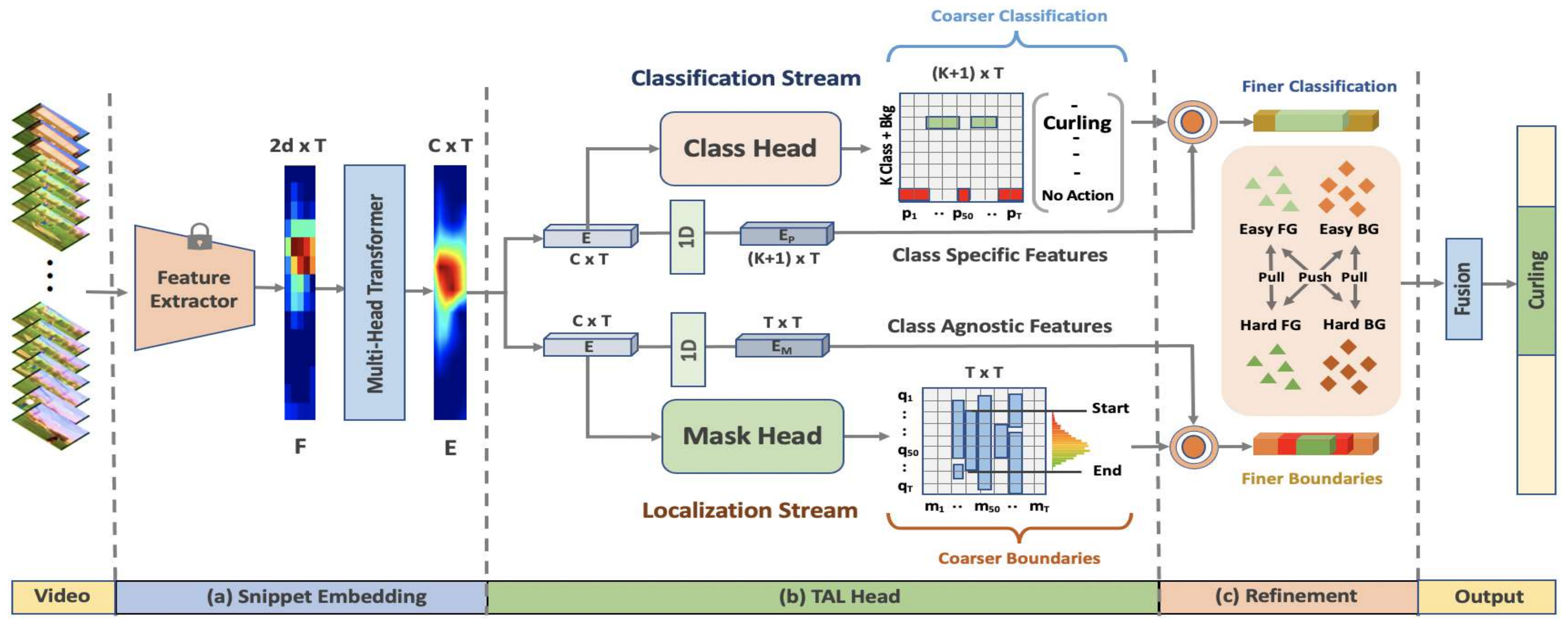
Background: Temporal action detection (TAD) aims to identify the temporal interval (i.e., the start and end points) and the class label of all action instances in an untrimmed video.

Motivation: Most SOTA TAD methods rely on large number of videos with exhaustive segment-level annotations. Existing methods adopt an intuitive strategy of combining an existing TAD models and a SSL method. This strategy is sub-optimal and prone to an error propagation problem. As illustrated in Fig. 1(a), this is because existing TAD models adopt a sequential localization (e.g., proposal generation) and classification design. When extended to SSL setting, the localization errors, can be easily propagated to the classification module leading to accumulated errors in class prediction in low-data setting.

Contributions: (1) To solve the localization error propagation problem suffered by existing SS-TAD methods, we propose a Proposal-Free Temporal Masking (SPOT) model with a new parallel classification and localization architecture; (2) We further design a novel pretext task for model pre-training and a boundary refinement algorithm; (3) SOTA performance on ActivityNet and THUMOS.

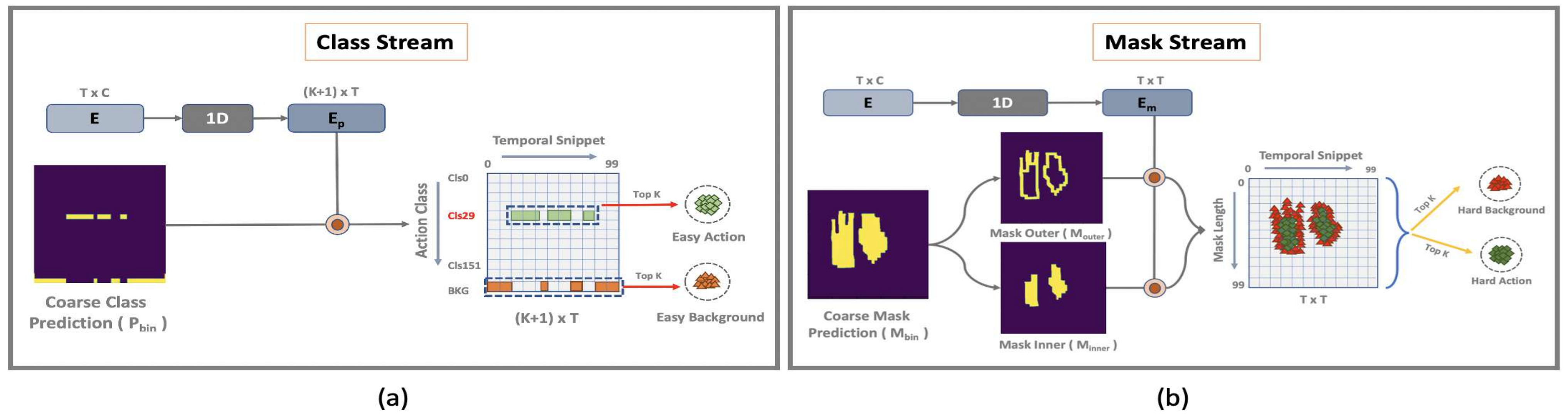


Model Architecture



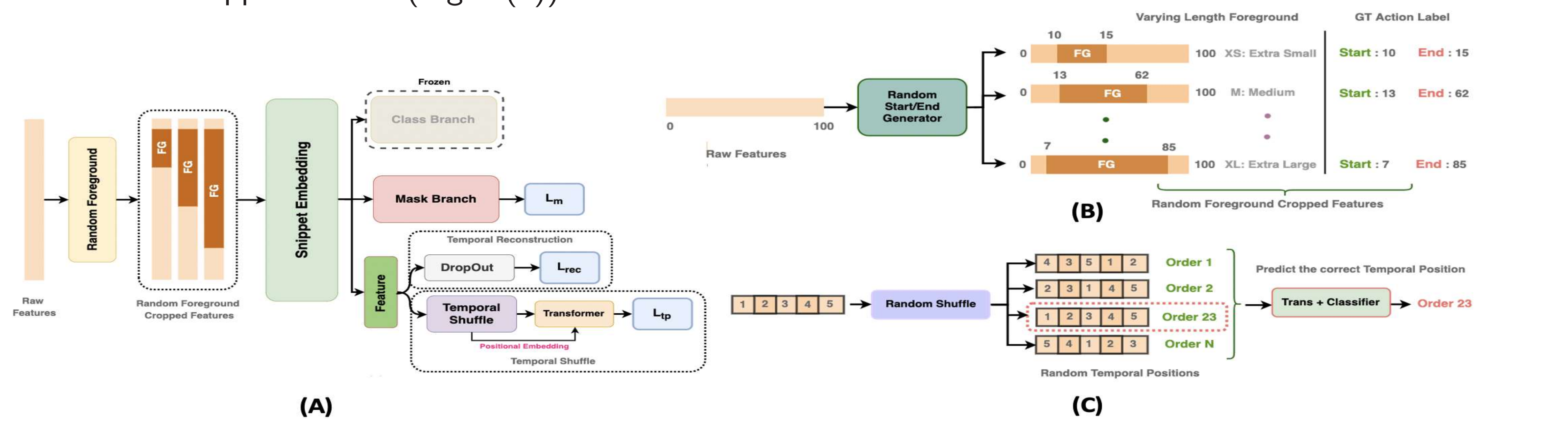
Contrastive Boundary Refinement

TAD methods typically struggle at estimating accurately the boundary between foreground and background segments. We refine them by obtaining hard and easy examples from the action boundary based on the structure of Class Branch (Fig 3(a)) and Mask Branch (Fig. 3a(b)) and calculate a infoNCE contrastive objective between them.



Two Stage Training: Pretrain then Finetune

Stage-I: Unsupervised Pretraining: We introduce a pretext task based on a novel notion of random foreground. With a randomly masked feature sequence, our pretext task aims to predict jointly (**1**) the temporal mask with the start s and end e (Fig. 4(b)), (**2**) the temporal position of each snippet after shuffling (Fig. 4(c)), and (**3**) the reconstruction of snippet feature (Fig. 4(a)).



Stage-II: Semi-Supervised Finetuning: We implement temporal mask semisupervised learning following the pseudo label paradigm. Concretely, we alternate between predicting and applying pseudo labels, starting by using the labeled samples alone. Note that the temporal ordering loss term is not used during fine-tuning as it gives performance drop.

Main Results

Results on ActivityNetv1.3 and THUMOS14

Labels	Methods	ActivityNet				THUMOS					
		0.5	0.75	0.95	Avg	0.3	0.4	0.5	0.6	0.7	Avg
60%	BMN*	47.6	31.7	7.5	31.5	50.8	45.9	34.8	23.7	16.3	34.3
	Mean Teacher+BMN	48.0	32.1	7.4	31.9	53.5	45.0	36.9	27.4	19.0	35.8
	FixMatch+BMN	48.7	32.9	7.7	32.8	53.8	46.2	37.8	28.7	19.5	36.9
	SSP	49.8	34.5	7.0	33.5	53.2	46.8	39.3	29.7	19.8	37.8
	SSTAP	50.1	34.9	7.4	34.0	56.4	49.5	41.0	30.9	21.6	39.9
	SPOT (Ours)	**52.8**	**35.0**	**8.1**	**35.2**	**58.9**	**50.1**	**42.3**	**33.5**	**22.9**	**41.5**
10%	BMN*	35.4	26.4	8.0	25.8	38.3	28.3	18.8	11.4	5.6	20.5
	Mean Teacher+BMN	36.0	27.2	7.4	26.6	41.2	32.1	23.1	15.0	7.0	23.7
	FixMatch+BMN	36.8	27.9	8.0	26.9	42.0	32.8	23.0	15.9	8.5	24.3
	SSP	38.9	28.7	8.4	27.6	44.2	34.1	24.6	16.9	9.3	25.8
	SSTAP	40.7	29.6	**9.0**	28.2	45.6	35.2	26.3	17.5	10.7	27.0
	SPOT (Ours)	**49.9**	**31.1**	8.3	**32.1**	**49.4**	**40.4**	**31.5**	**22.9**	**12.4**	**31.3**

Ablation Studies

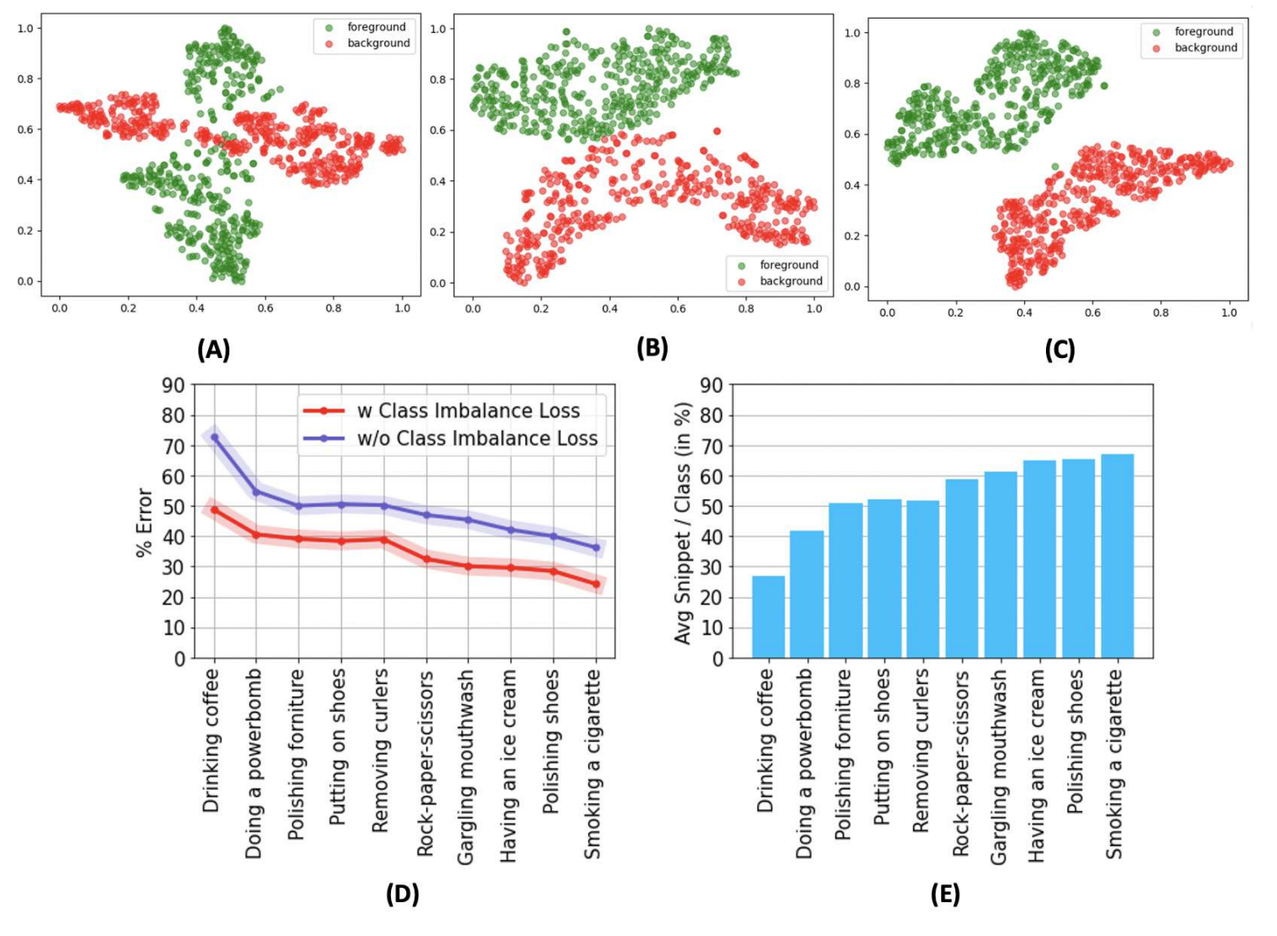
Analysis of localization error propagation

Metric	mAP	
	0.5	Avg
BMN [?] + MLP		
GT proposals	55.7	45.3
Pseudo proposals	32.4	23.6
SPOT		
GT masks	59.2	47.0
Pseudo masks	49.9	32.1

SPOT model w/ and w/o unlabeled data.

Labels	SSL Modules		mAP	
	Pre-train	L_c	0.5	Avg
10%	✓	✓	**49.9**	**32.1**
	✗	✓	46.2	30.5
	✗	✗	44.5	28.3
60%	✓	✓	**52.8**	**35.2**
	✗	✓	52.1	34.9
	✗	✗	51.2	34.0

Effect of SPOT in Feature Space.


(A)
(B)
(C)
w Class Imbalance Loss
w/o Class Imbalance Loss
% Error
Avg Snippet / Class (in %)
(D)
(E)