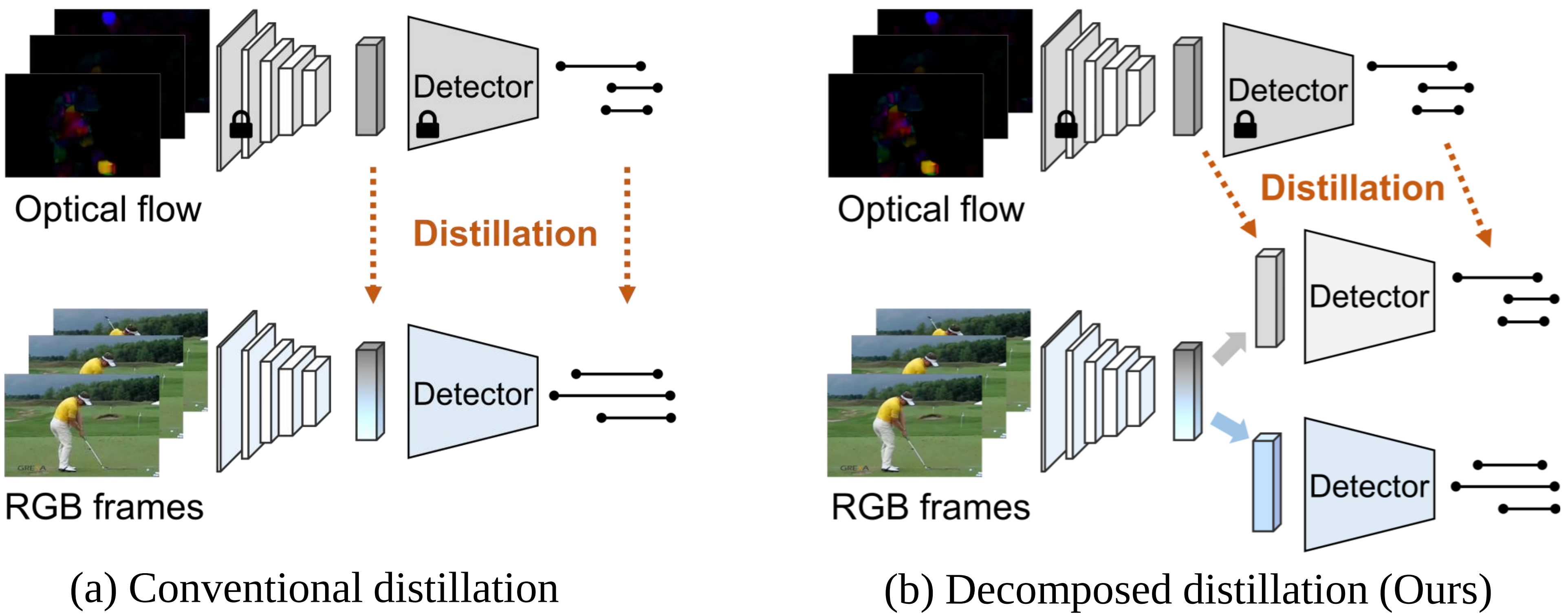
TUE-AM-227

Decomposed Cross-modal Distillation for RGB-based Temporal Action Detection

Pilhyeon Lee   Taeoh Kim   Minho Shim   Dongyoon Wee   Hyeran Byun

Summary

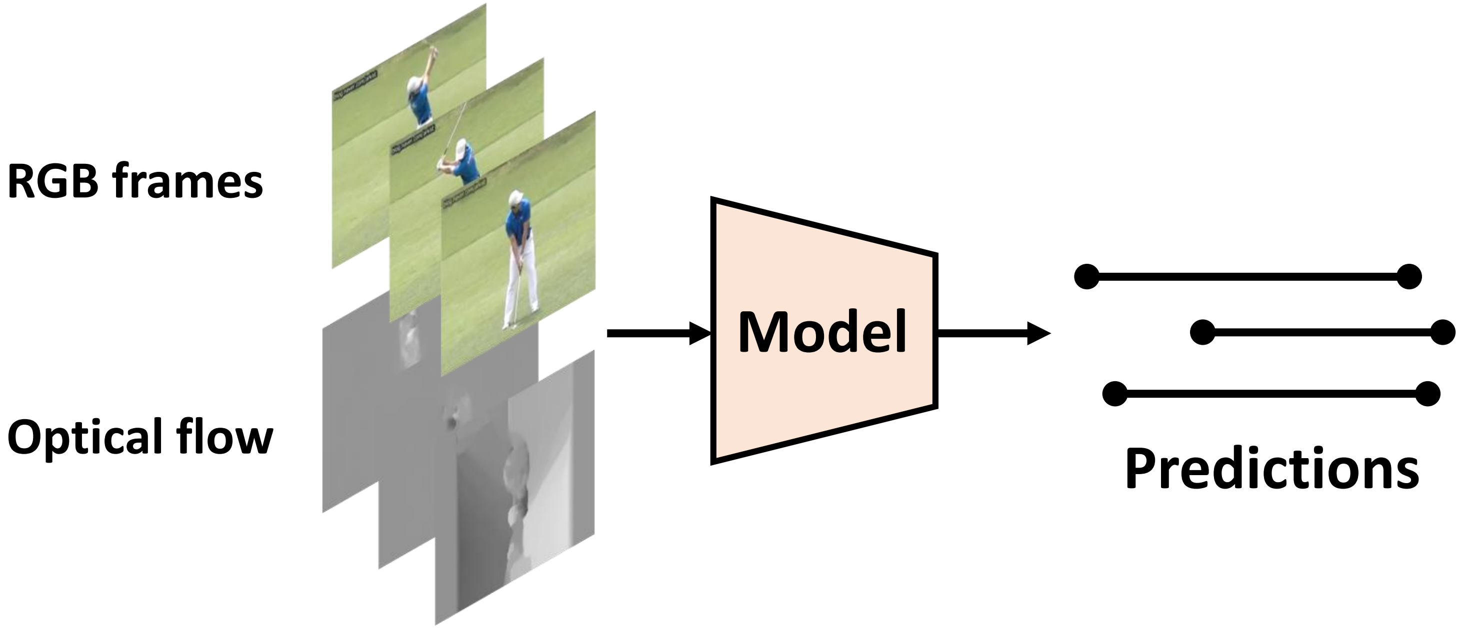


We propose to learn appearance and motion features in a **decomposed** way to better exploit the multimodal complementarity.

Temporal action detection

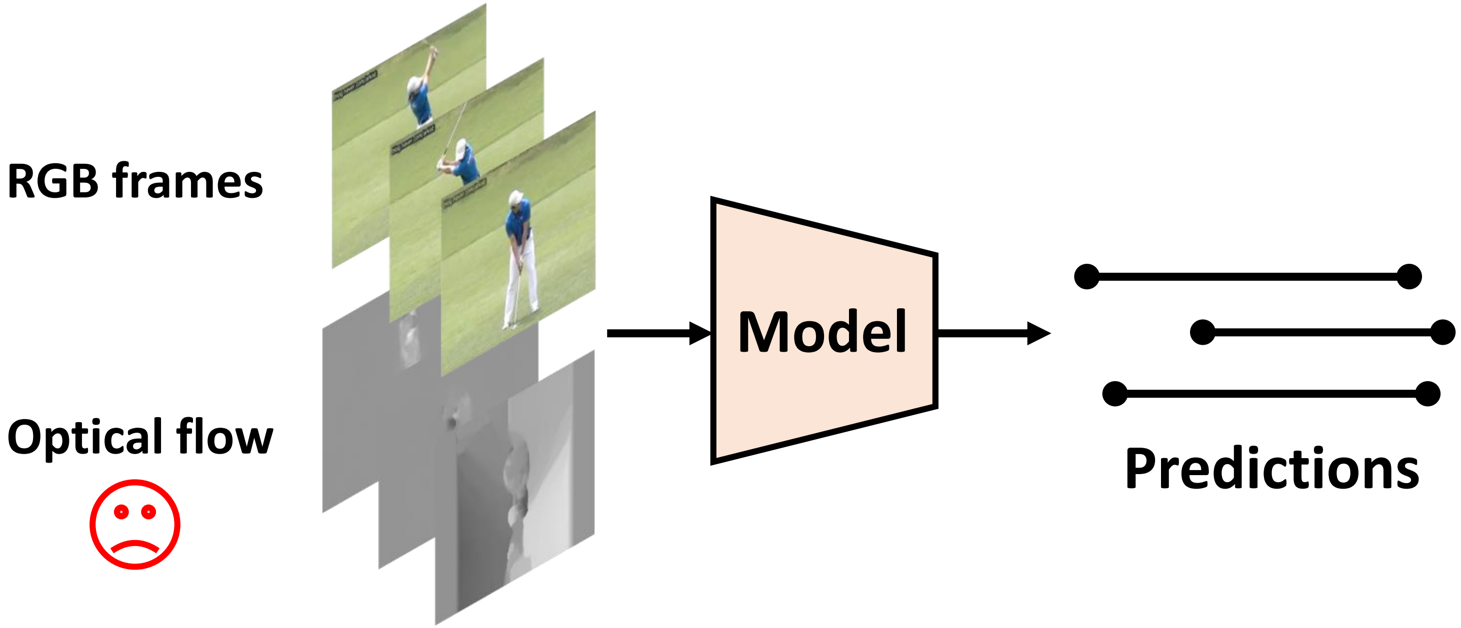
Goal: to predict the *temporal intervals* and *classes* of action instances.

Temporal action detection



Existing approaches commonly leverage two modalities,
i.e., RGB and optical flow, for precise action detection.

Heavy cost of optical flow



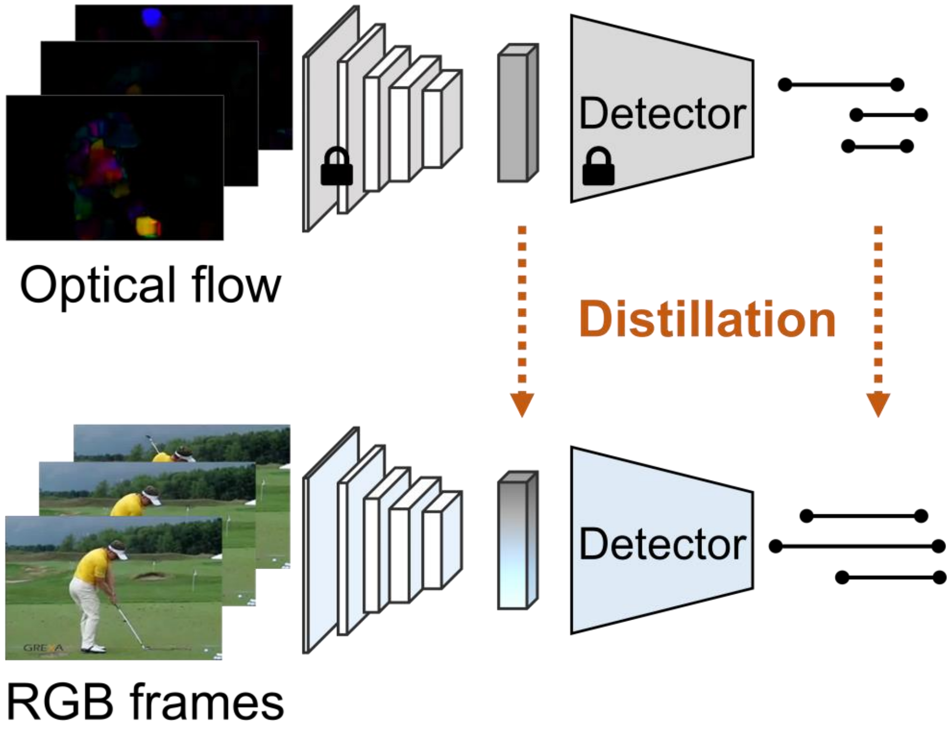
Optical flow is computationally expensive,
e.g., TV-L^1 requires 3.8 minutes for a 1-min 224 × 224 video of 30 fps.

Reliance of action detectors on optical flow

Framework	Method	Average mAP (%) RGB+OF	RGB	Δ
Anchor-based	G-TAD [74]	41.5	26.9	−14.6
Anchor-free	AFSD [34]	52.4	43.3	−9.1
	Actionformer [80]	62.2	55.5	−6.7
DETR-like	TadTR [42]	56.7	46.0	−10.7
Proposal-free	TAGS [47]	52.8	47.9	−4.9

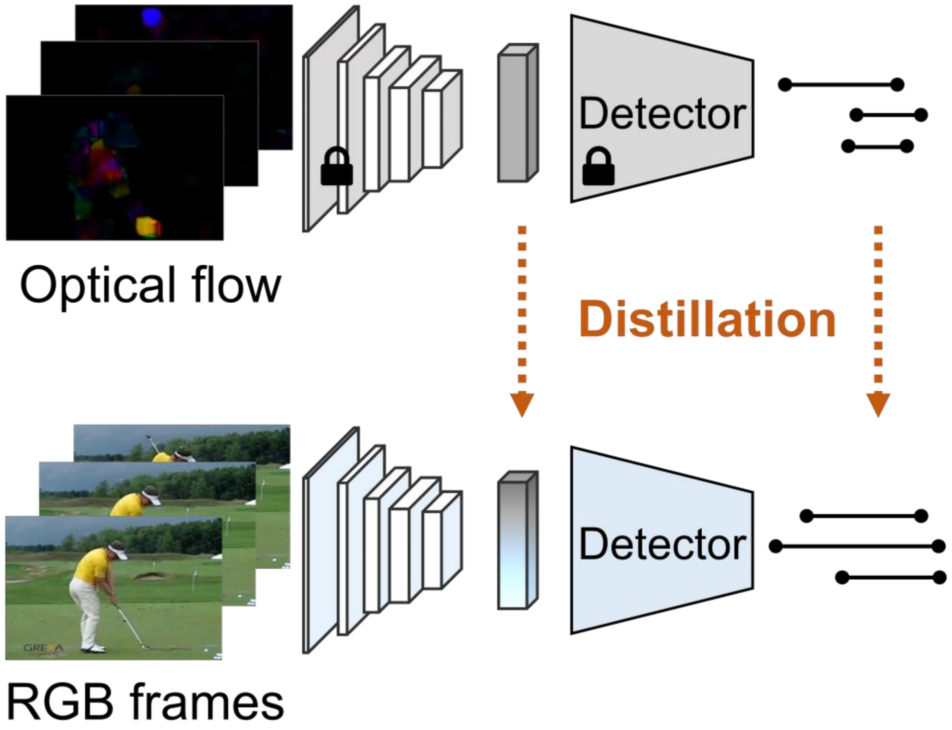
Existing temporal action detectors heavily rely on optical flow;
they show sharp performance drops in the absence of optical flow.

Cross-modal knowledge distillation



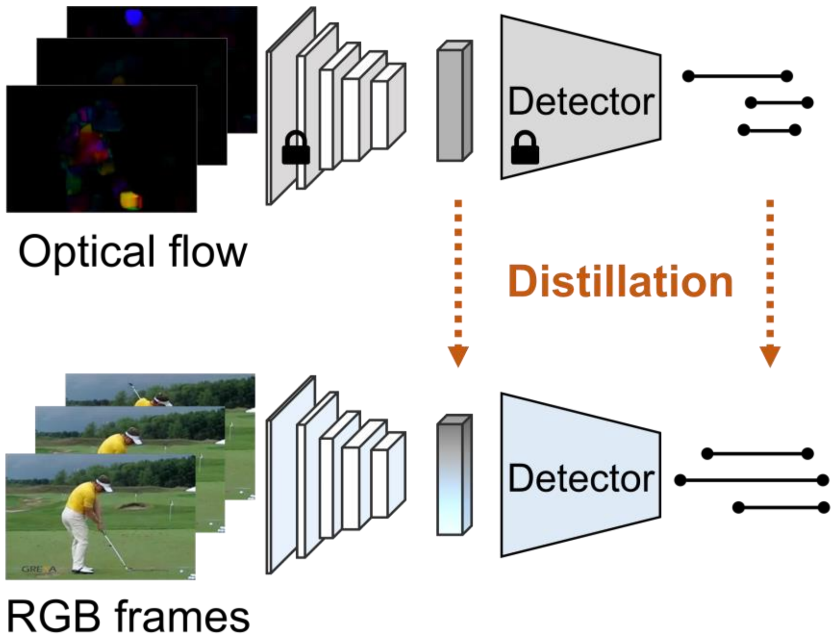
Cross-modal knowledge distillation transfers motion knowledge the RGB-based model, enhancing its performance.

Cross-modal knowledge distillation

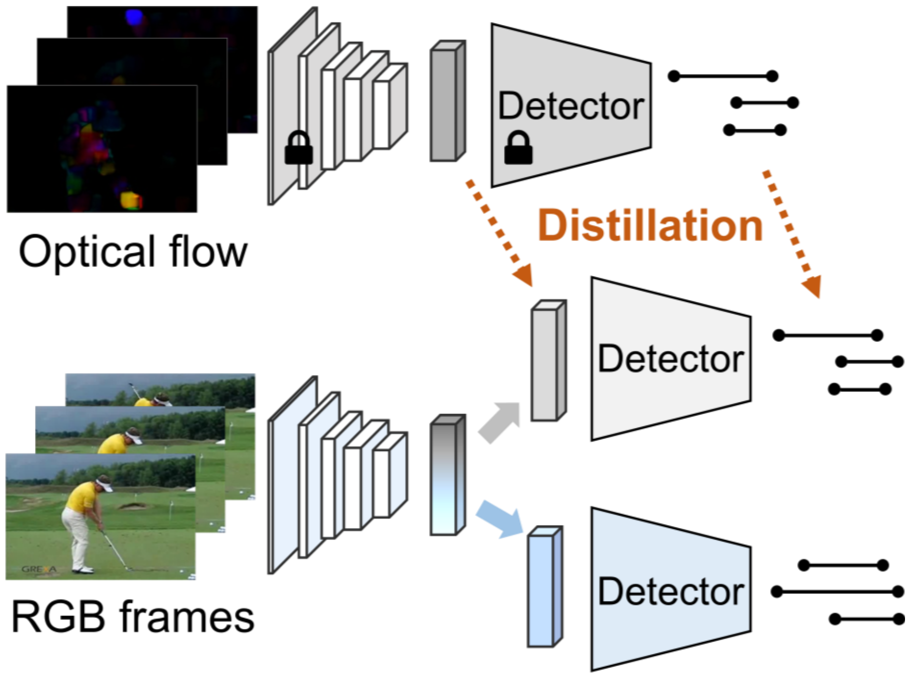


Conventional distillation leads to entangled multimodal representations, making it challenging to balance between two modalities.

Decomposed cross-modal knowledge distillation



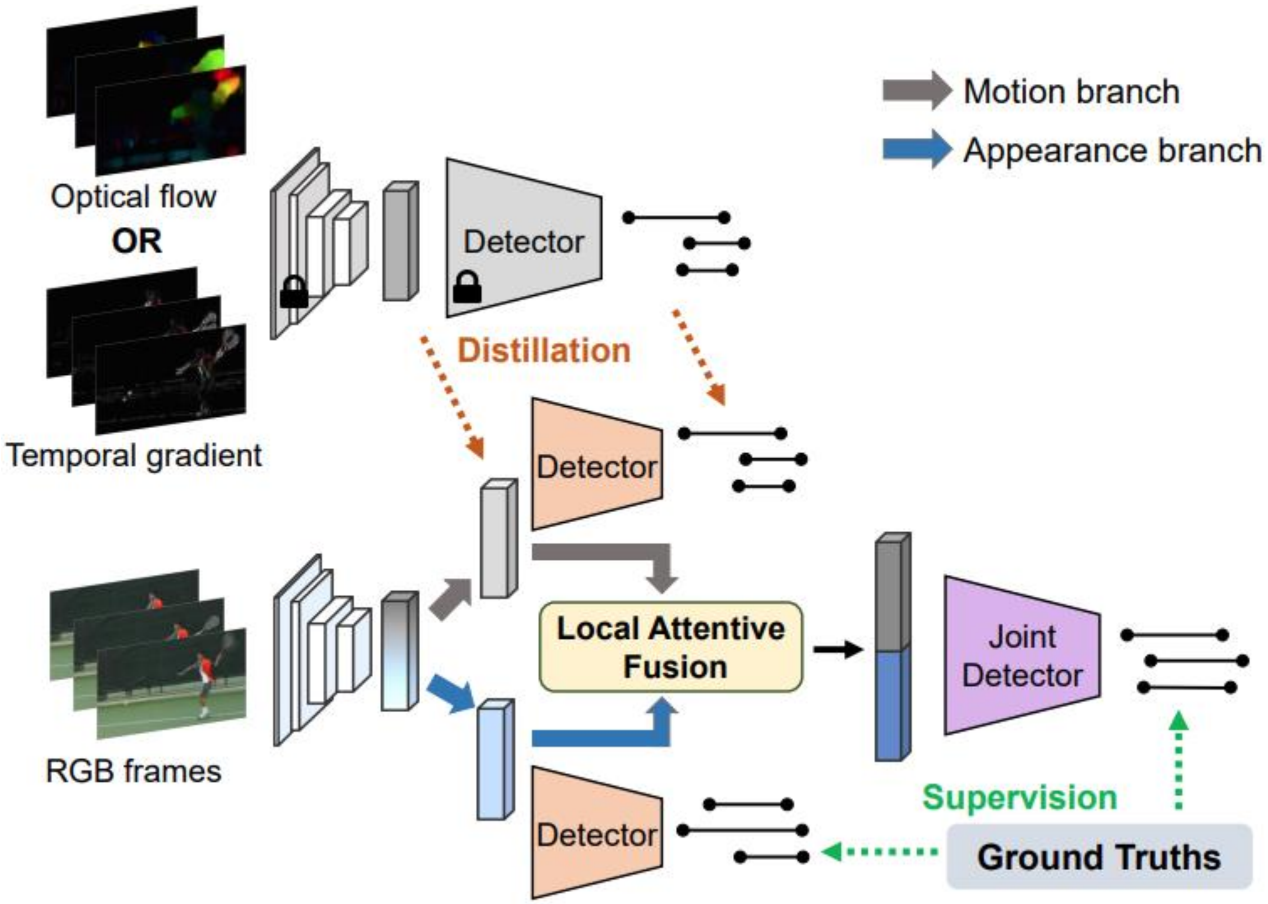
(a) Conventional distillation



(b) Decomposed distillation (Ours)

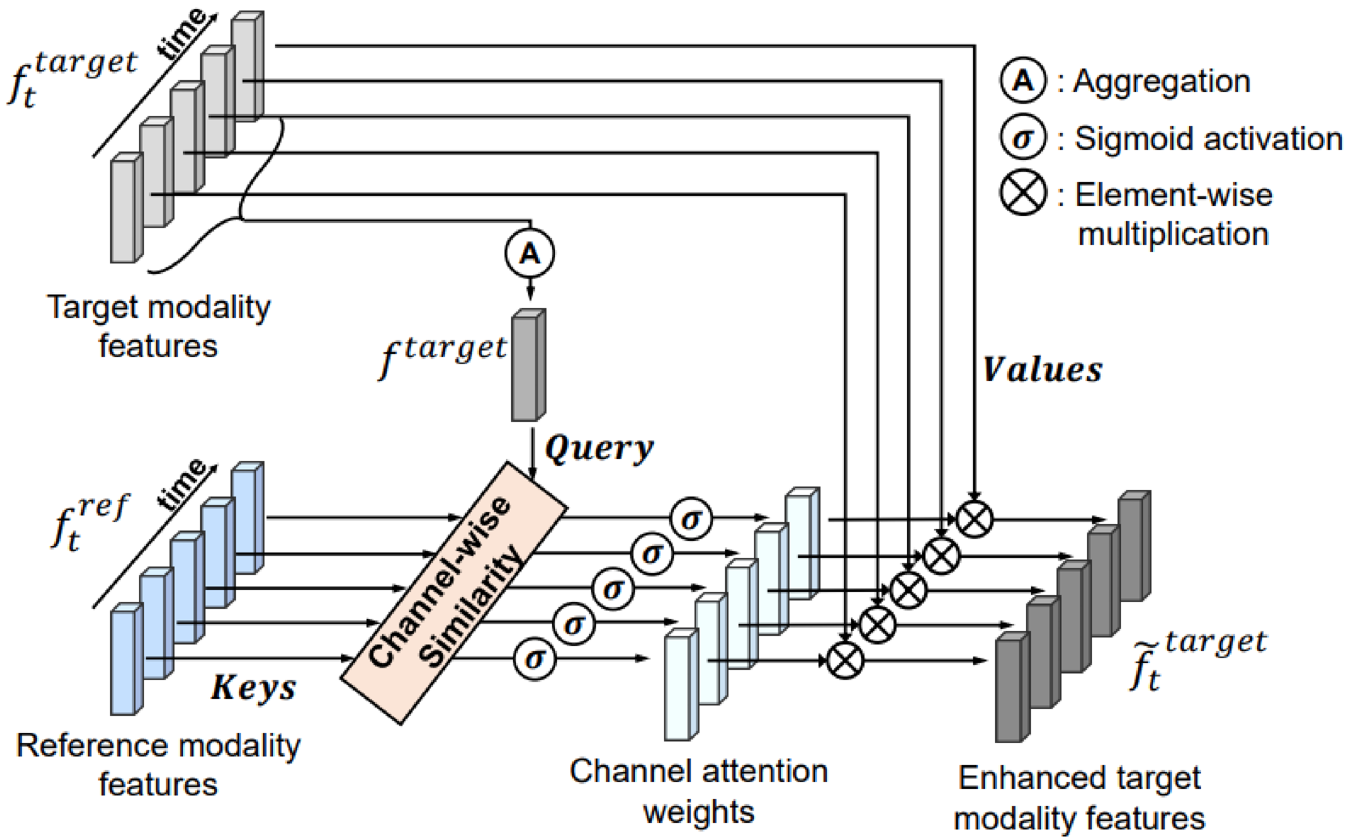
We propose to learn appearance and motion features in a **decomposed** way to better exploit the multimodal complementarity.

Method



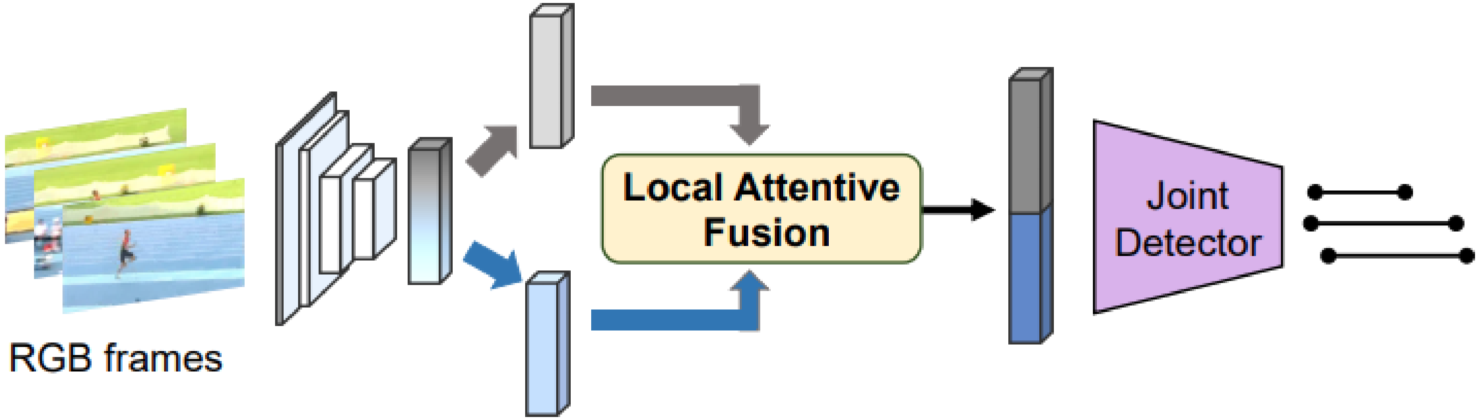
We design a dual-branch architecture with a shared head and conflicting training objectives for explicit decomposition of multimodal information.

Method



The local attentive fusion enables effective multimodal information fusion while bypassing the feature over-smoothing issue.

Method



At inference time, our model can perform multimodal prediction given only RGB frames as input.

Experiments

distillation conven.	distillation decomp.	local attn.	mAP@IoU (%) 0.3	0.4	0.5	0.6	0.7	AVG
✗	✗	✗	62.3	55.2	46.2	33.8	20.4	43.6
✓			62.5	55.7	47.3	35.1	21.8	44.5
	✓		63.3	56.2	47.9	36.1	22.9	45.2
	✓	✓	64.4	58.0	49.0	37.5	24.1	46.6

Ablative studies verify the effectiveness of the each proposed component.

Experiments

Fusion	mAP@IoU (%) 0.3	0.4	0.5	0.6	0.7	AVG
concat.	63.3	56.2	47.9	36.1	22.9	45.2
sum.	62.6	56.1	47.5	36.1	23.0	45.1
self-attn.	63.8	56.3	46.7	34.2	21.9	44.6
cross-attn.	63.1	54.5	46.4	35.4	21.7	44.2
diff.-attn.	61.8	54.8	46.3	32.6	21.0	43.3
local attn. (Ours)	64.4	58.0	49.0	37.5	24.1	46.6

The local attentive fusion brings the largest performance gains compared to other fusion methods.

Experiments

Backbone	Distill.	mAP@IoU (%) 0.3	0.4	0.5	0.6	0.7	AVG
TSM18 [29]	✗	62.3	55.2	46.2	33.8	20.4	43.6
	TG	64.4	58.0	49.0	37.5	24.1	46.6 (+3.0)
	OF	65.3	59.5	50.9	39.6	25.5	48.2 (+4.6)
TSM50 [29]	✗	65.0	59.2	50.0	38.2	25.0	47.5
	TG	68.1	61.8	52.4	41.7	27.5	50.3 (+2.8)
	OF	66.5	62.3	55.3	44.5	32.9	52.3 (+4.8)
I3D [6]	✗	53.8	47.0	38.6	30.0	19.9	37.9
	TG	57.6	51.4	42.5	32.9	22.1	41.3 (+3.4)
	OF	57.7	52.1	44.6	34.9	24.0	42.6 (+4.7)
Slowfast50 [15]	✗	67.4	62.9	56.8	46.8	35.0	53.8
	TG	68.9	64.1	58.1	48.2	35.6	55.0 (+1.2)
	OF	70.5	65.8	59.2	50.1	38.2	56.8 (+3.0)

Head	Distill.	mAP@IoU (%) 0.3	0.4	0.5	0.6	0.7	AVG
G-TAD [67]	✗	51.4	44.7	36.0	26.4	16.8	35.1
	TG	54.8	48.9	38.1	28.0	18.1	37.6 (+2.5)
	OF	55.3	49.4	39.2	30.6	19.7	38.8 (+3.6)
TadTR [36]	✗	62.8	56.7	47.5	37.3	25.5	46.0
	TG	63.8	57.4	49.9	39.2	26.9	47.4 (+1.4)
	OF	64.1	58.3	51.2	40.9	28.8	48.7 (+2.7)
Actionformer [73]	✗	62.3	55.2	46.2	33.8	20.4	43.6
	TG	64.4	58.0	49.0	37.5	24.1	46.6 (+3.0)
	OF	65.3	59.5	50.9	39.6	25.5	48.2 (+4.6)

The proposed method is generalizable to various backbones and action detection heads.

Experiments

Method	Venue	OF	THUMOS'14 0.3	0.4	0.5	0.6	0.7	AVG	ActivityNet 0.5	0.75	0.95	AVG
TAL-Net [7]	CVPR'18	✓	53.2	48.5	42.8	33.8	20.8	39.8	38.23	18.30	1.30	20.22
BSN [31]	ECCV'18	✓	53.5	45.0	36.9	28.4	20.0	-	46.45	29.96	8.02	30.03
BMN [30]	ICCV'19	✓	56.0	47.4	38.8	29.7	20.5	38.5	50.07	34.70	8.29	33.85
P-GCN [72]	ICCV'19	✓	63.6	57.8	49.1	-	-	-	48.26	33.16	3.27	31.11
G-TAD [67]	CVPR'20	✓	54.5	47.6	40.2	30.8	23.4	39.3	50.36	34.60	9.02	34.09
BC-GNN [2]	ECCV'20	✓	57.1	49.1	40.4	31.2	23.1	40.2	50.56	34.75	9.37	34.26
BU-MR [77]	ECCV'20	✓	53.9	50.7	45.4	38.0	28.5	43.3	43.47	33.91	9.21	30.12
AFSD [28]	CVPR'21	✓	67.3	62.4	55.5	43.7	31.1	52.0	52.38	35.27	6.47	34.39
MUSES [35]	CVPR'21	✓	68.9	64.0	56.9	46.3	31.0	53.4	50.02	34.97	6.57	33.99
RTD-Net [53]	ICCV'21	✓	68.3	62.3	51.9	38.8	23.7	49.0	47.21	30.68	8.61	30.83
VSGN [75]	ICCV'21	✓	66.7	60.4	52.4	41.0	30.4	50.2	52.38	36.01	8.37	35.07
RCL [57]	CVPR'22	✓	70.1	62.3	52.9	42.7	30.7	51.7	55.15	39.02	8.27	37.65
RefactorNet [61]	CVPR'22	✓	70.7	65.4	58.6	47.0	32.1	54.8	56.60	40.70	7.50	38.60
TAGS [41]	ECCV'22	✓	68.6	63.8	57.0	46.3	31.8	52.8	56.30	36.80	9.60	36.50
ReAct [47]	ECCV'22	✓	69.2	65.0	57.1	47.8	35.6	55.0	49.60	33.00	8.60	32.60
Actionformer [73]	ECCV'22	✓	82.1	77.8	71.0	59.4	43.9	66.8	53.50	36.20	8.20	35.60
CDC [49]	CVPR'17	✗	40.1	29.4	23.3	13.1	7.9	22.8	45.30	26.00	0.20	23.80
GTAN [38]	CVPR'19	✗	57.8	47.2	38.8	-	-	-	52.61	34.14	8.91	34.31
G-TAD* [67]	CVPR'20	✗	52.5	45.9	37.6	28.5	19.1	36.7	49.22	34.55	4.74	33.17
AFSD* [28]	CVPR'21	✗	57.7	52.8	45.4	34.9	22.0	43.6	-	-	-	32.90
TadTR* [36]	TIP'22	✗	59.6	54.5	47.0	37.8	26.5	45.1	49.56	35.24	9.93	34.35
E2E-TAD [34]	CVPR'22	✗	69.4	64.3	56.0	46.4	34.9	54.2	50.47	35.99	10.83	35.10
TAGS† [41]	ECCV'22	✗	59.8	57.2	50.7	42.6	29.1	47.9	54.44	34.95	8.71	34.95
Actionformer† [73]	ECCV'22	✗	69.8	66.0	58.7	48.3	34.6	55.5	53.21	35.15	8.03	34.94
Ours	-	✗	70.5	65.8	59.2	50.1	38.2	56.8	53.73	35.87	8.61	35.58

Our method achieves a new state-of-the-art among RGB-based action detectors, closing the gap with two-stream approaches.

Conclusion

- We introduced a novel cross-modal distillation pipeline that learns multimodal information in a decomposed way.
- Our method generalizes well to different backbones and action detection heads, showing consistent improvements.
- Our approach is abstract and can be applied to various multimodal tasks that require multimodal complementarity.

Thank you!

Contact: lph1114@yonsei.ac.kr