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

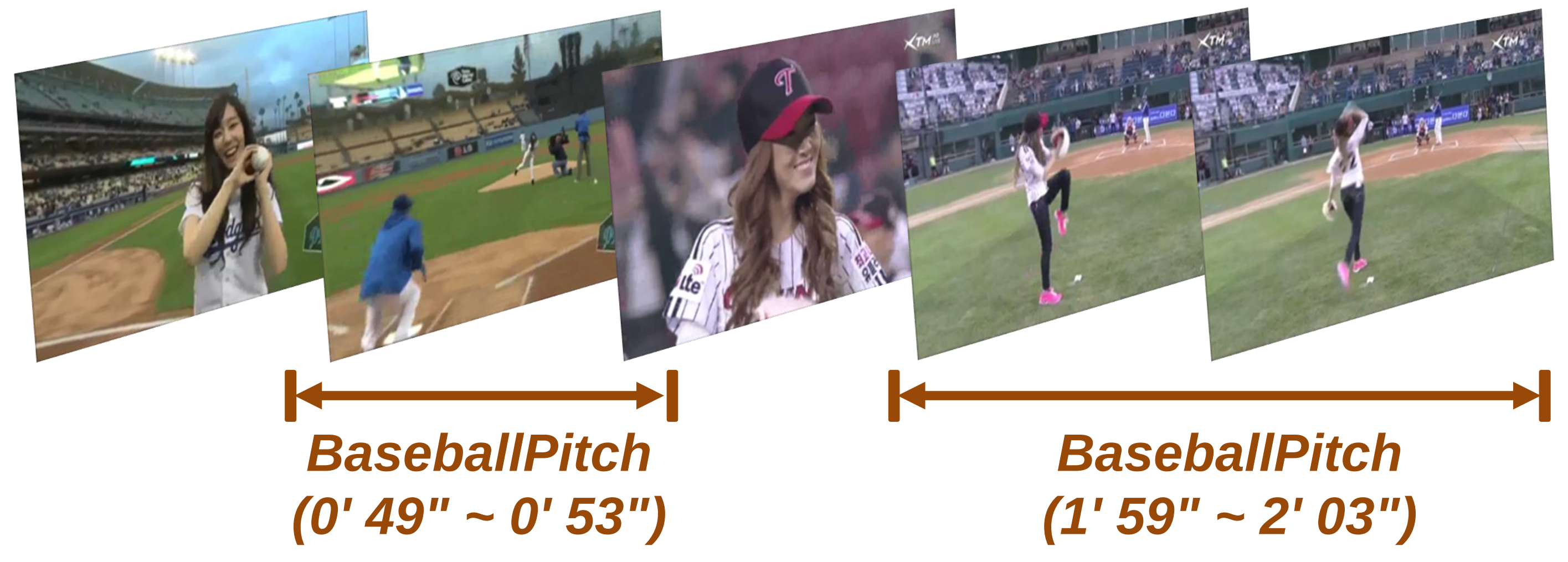
Pilhyeon Lee Taeoh Kim Minho Shim Dongyoon Wee Hyeran Byun

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Problem

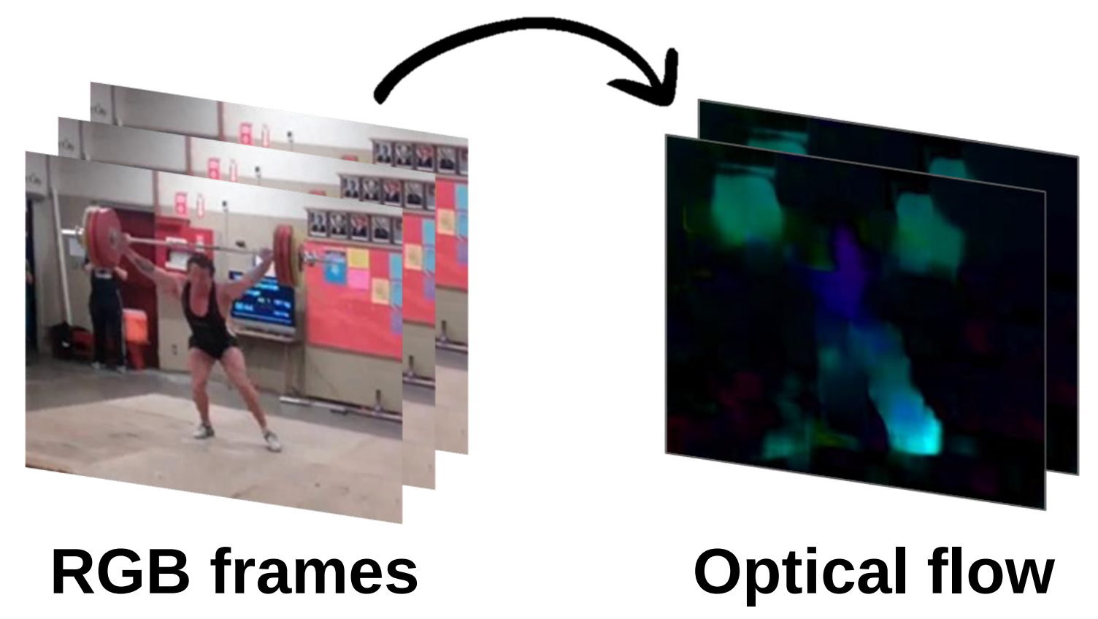
Temporal action detection (or localization)

The goal is to predict action intervals and their classes.



Introduction

- To date, a variety of temporal action detection models have shown promising performance based on two-stream inputs.



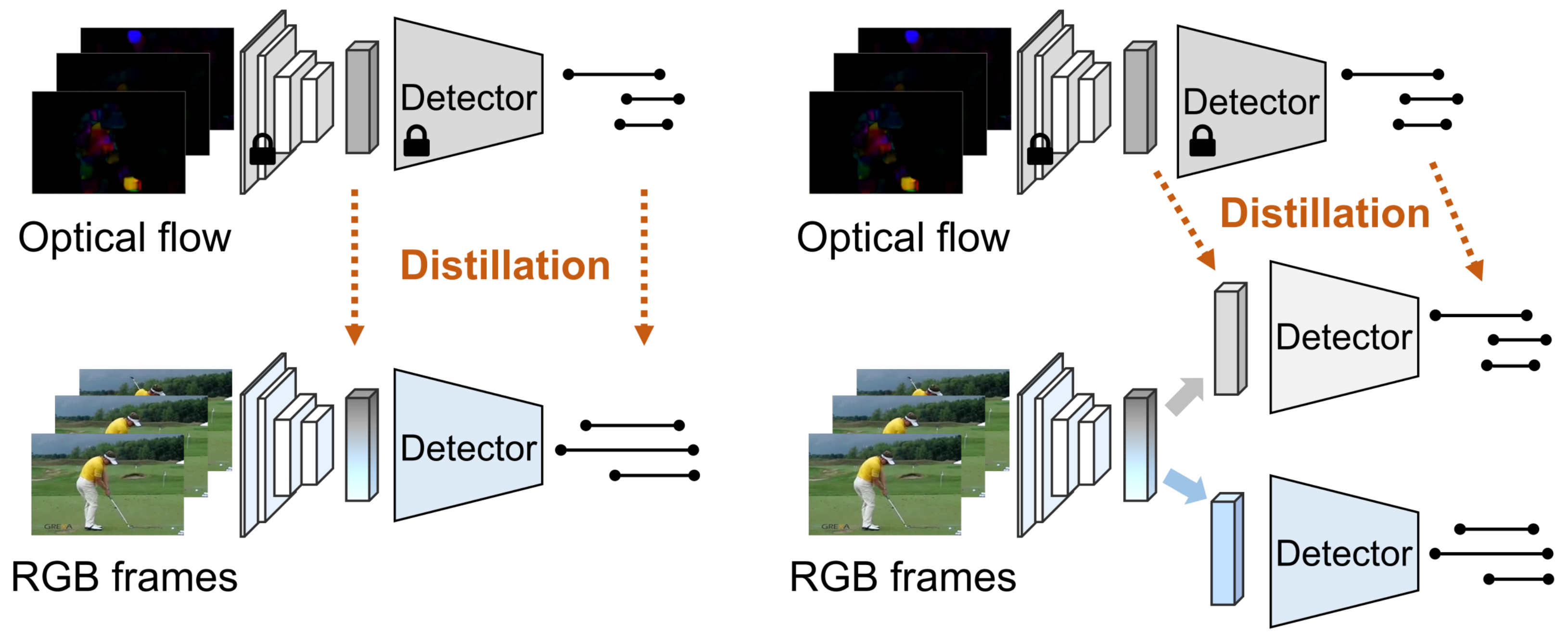
- However, they all heavily rely on computationally expensive optical flow for performance, regardless of framework types.

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

- **How costly is optical flow?** E.g., TV-L^1 takes 1.8 minutes to process a 1-min 224 × 224 video of 30 fps on a GPU.
- To bypass the cost, we aim to build a strong RGB-based action detector for both **efficient** and **accurate** prediction with a novel cross-modal knowledge distillation framework.

Decomposed Cross-modal Distillation

Contrary to conventional distillation where multimodal information is entangled, the proposed decomposed distillation framework encourages the model to learn it **in a decomposed way** to exploit better multimodal complementarity.

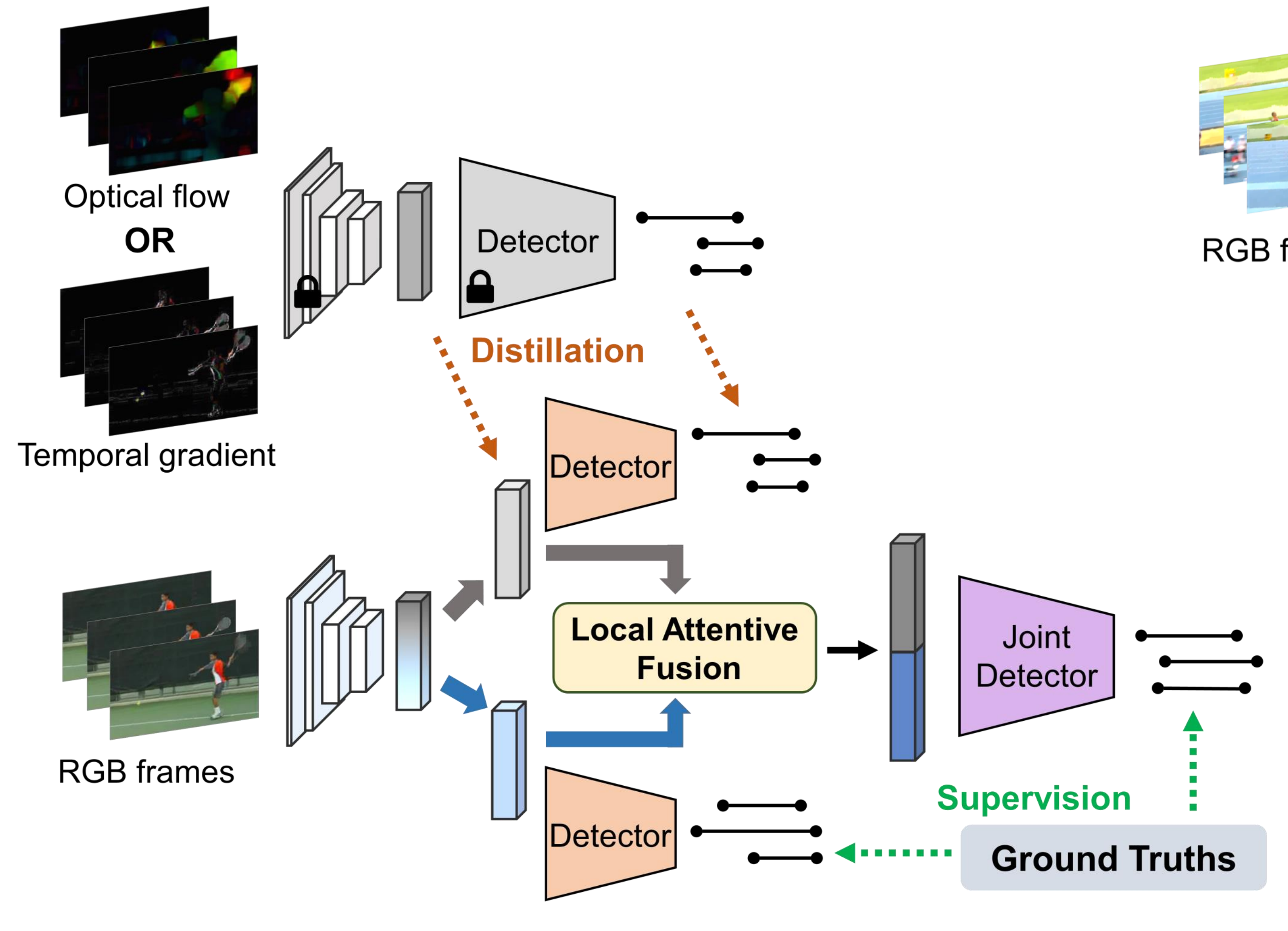


(a) Conventional distillation (b) Decomposed distillation (Ours)

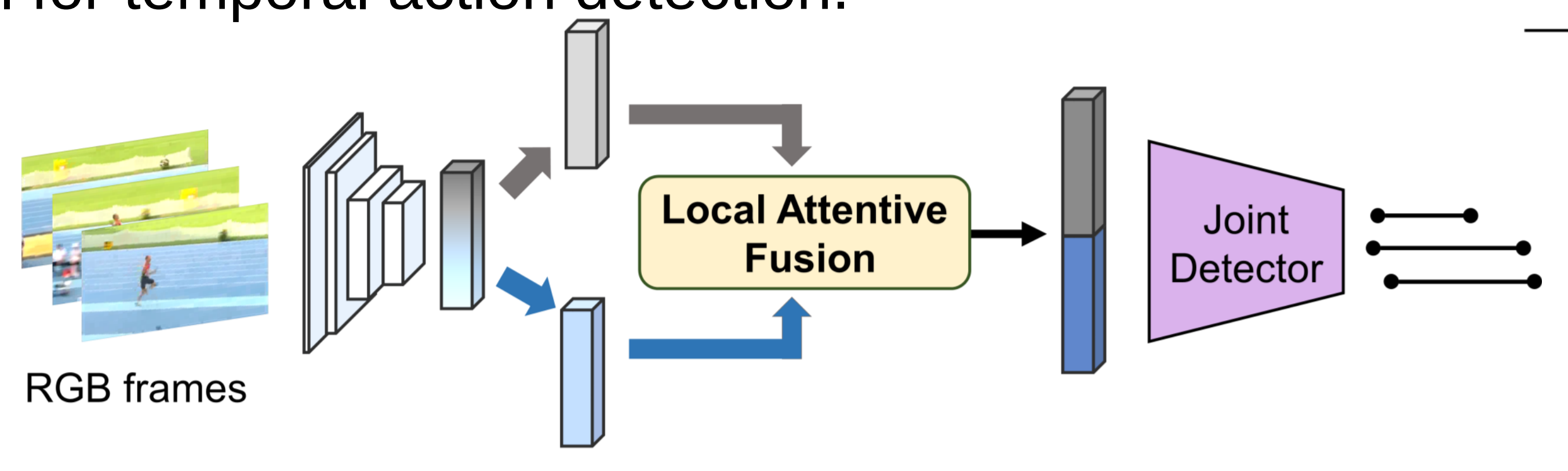
Architecture

Our model explicitly separates the motion and appearance features within a **dual-branch pipeline** where the two branches share the detection head but pursue conflicting training objectives.

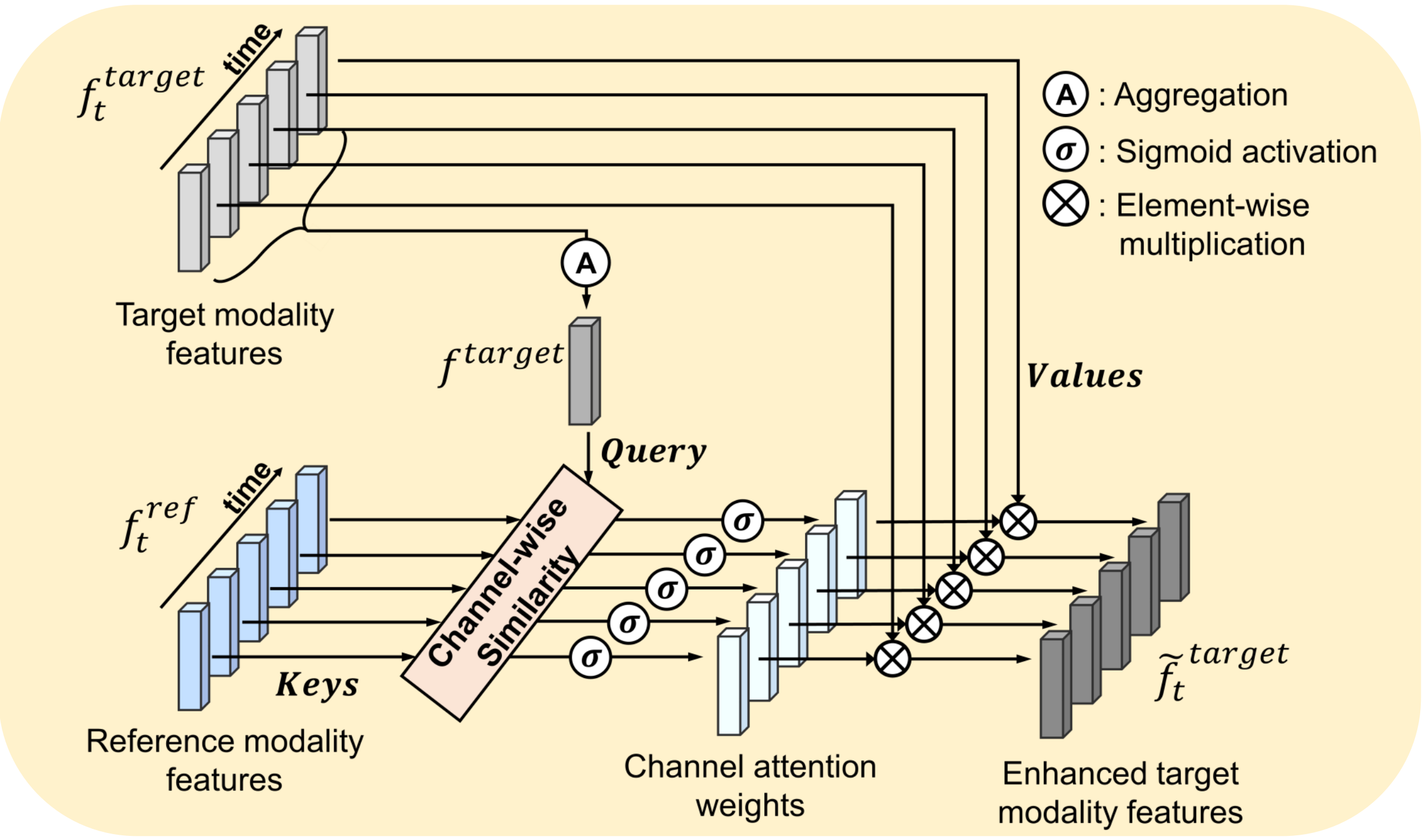
The proposed **local attentive fusion** enables effective multimodal information fusion while sustaining the local discriminability of features that is essential for temporal action detection.



(a) Training



(b) Inference



Experiments

The proposed method significantly enhances the RGB-based action detectors, while being generalizable to various backbones and detection heads.

Ablation studies

distillation conven.	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

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

Generalization tests

Backbone	Distill.	mAP@IoU (%) 0.3	0.4	0.5	0.6	0.7	AVG
TSM18 [35]	✗	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 [35]	✗	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)

(TG: temporal gradient, OF: Optical flow)

Head	Distill.	mAP@IoU (%) 0.3	0.4	0.5	0.6	0.7	AVG
G-TAD [74]	✗	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 [42]	✗	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 [80]	✗	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)

State-of-the-art comparisons

Method	Venue	OF	THUMOS'14 0.3	0.4	0.5	0.6	0.7	AVG	ActivityNet1.3 0.5	0.75	0.95	AVG
CDC [55]	CVPR'17	✗	40.1	29.4	23.3	13.1	7.9	22.8	45.30	26.00	0.20	23.80
GTAN [44]	CVPR'19	✗	57.8	47.2	38.8	-	-	-	52.61	34.14	8.91	34.31
G-TAD* [74]	CVPR'20	✗	52.5	45.9	37.6	28.5	19.1	36.7	49.22	34.55	4.74	33.17
AFSD* [34]	CVPR'21	✗	57.7	52.8	45.4	34.9	22.0	43.6	-	-	-	32.90
TadTR* [42]	TIP'22	✗	59.6	54.5	47.0	37.8	26.5	45.1	49.56	35.24	9.93	34.35
E2E-TAD [40]	CVPR'22	✗	69.4	64.3	56.0	46.4	34.9	54.2	50.47	35.99	10.83	35.10
TAGS† [47]	ECCV'22	✗	59.8	57.2	50.7	42.6	29.1	47.9	54.44	34.95	8.71	34.95
Actionformer† [80]	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

Qualitative results