Cross-modal Consensus Network for Weakly Supervised Temporal Action Localization

Fa–Ting Hong[1,2,3,6,^,#], Jia–Chang Feng[1,3,^], Dan Xu[4], Ying Shan[2], Wei-Shi Zheng[1,3,5,*]

[1] School of Computer Science and Engineering, Sun Yat-sen University

[2] Applied Research Center (ARC), PCG, Tencent

[3] Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China

[4] Hong Kong University of Science and Technology

[5] Peng Cheng Laboratory, [6] Pazhou Lab

^ Equal contributed.

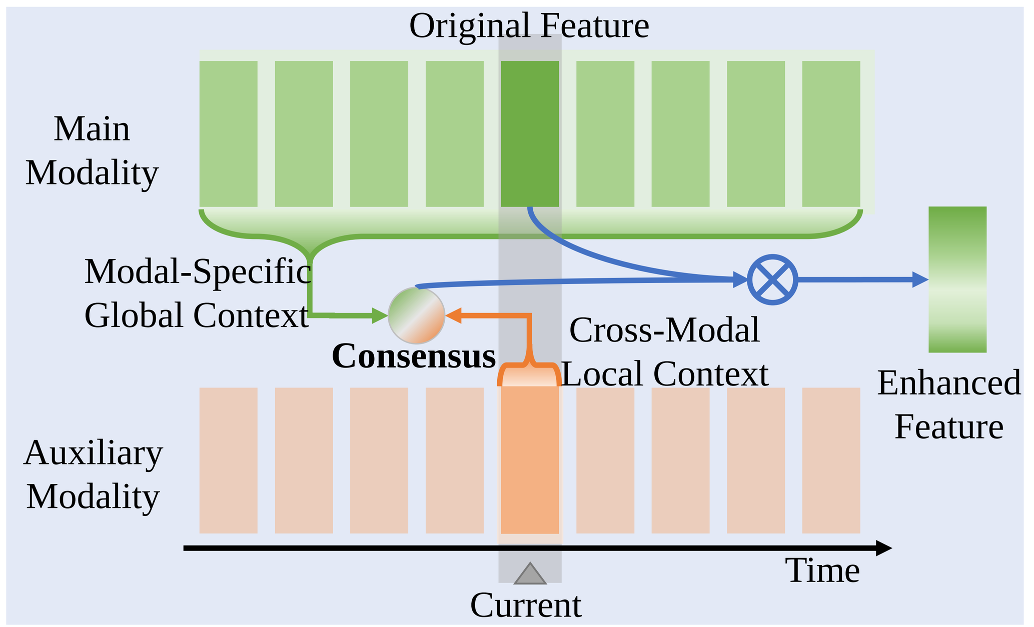
* Corresponding author.

\# Work done during internship at ARC

Project Page

ACM multimedia
Chengdu, China OCT 20-24 2021

Introduction



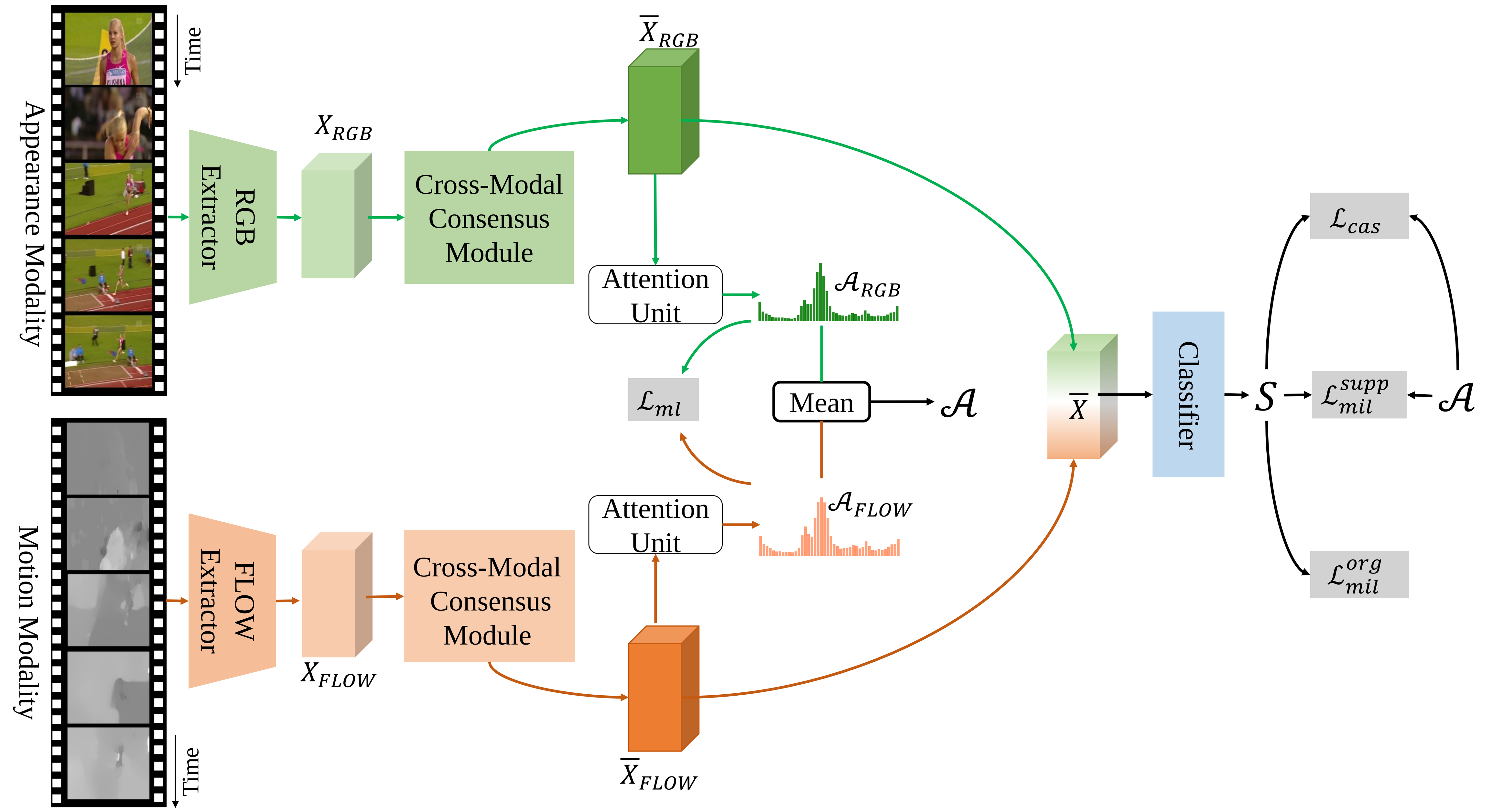
➢ Motivations:

- Inconsistency between feature encoder pretrained task and target task leads to task-irrelevant information in the extracted features.
- Previous works ignore the correlation between two types of modality but simply concatenation or score fusion.
- Inter-modality consistency can be further investigated.

➢ Contributions:

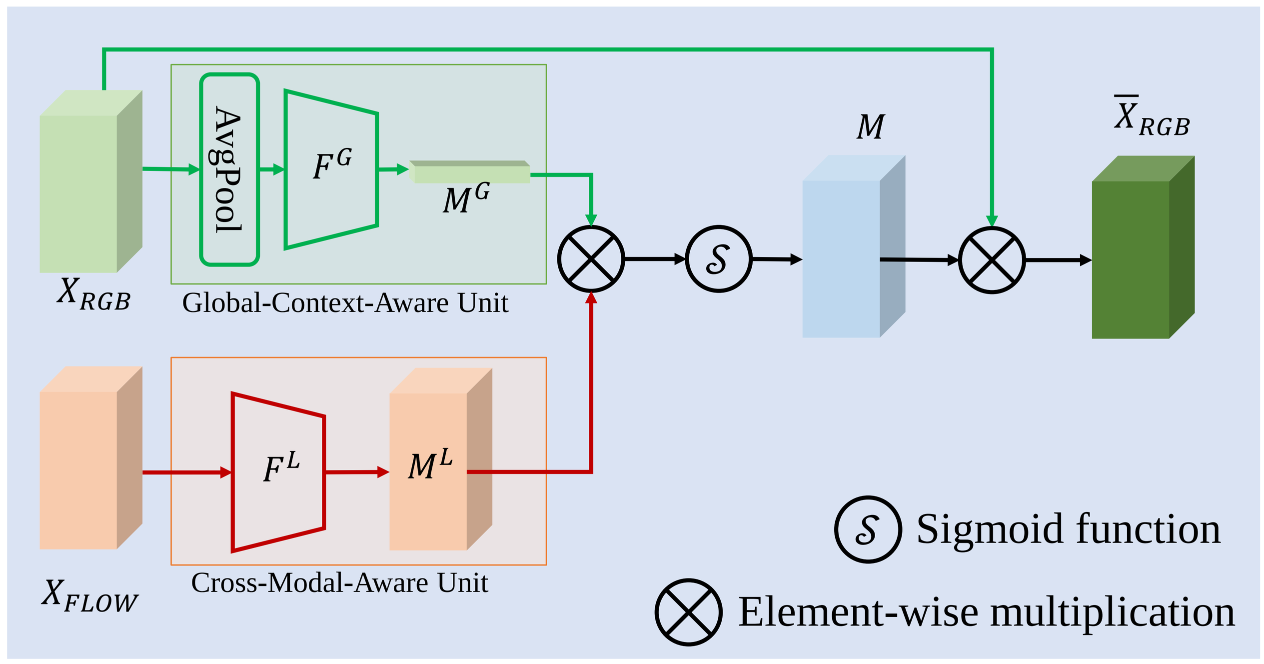
- This is the first work to investigate feature re-calibration in temporal action localization community.
- We propose to explore modal-wise consistency via mutual learning for temporal action localization.
- Our proposed CO2-Net outperform existing SOTA methods on two public benchmark without using any external information.

Overview



Methodology

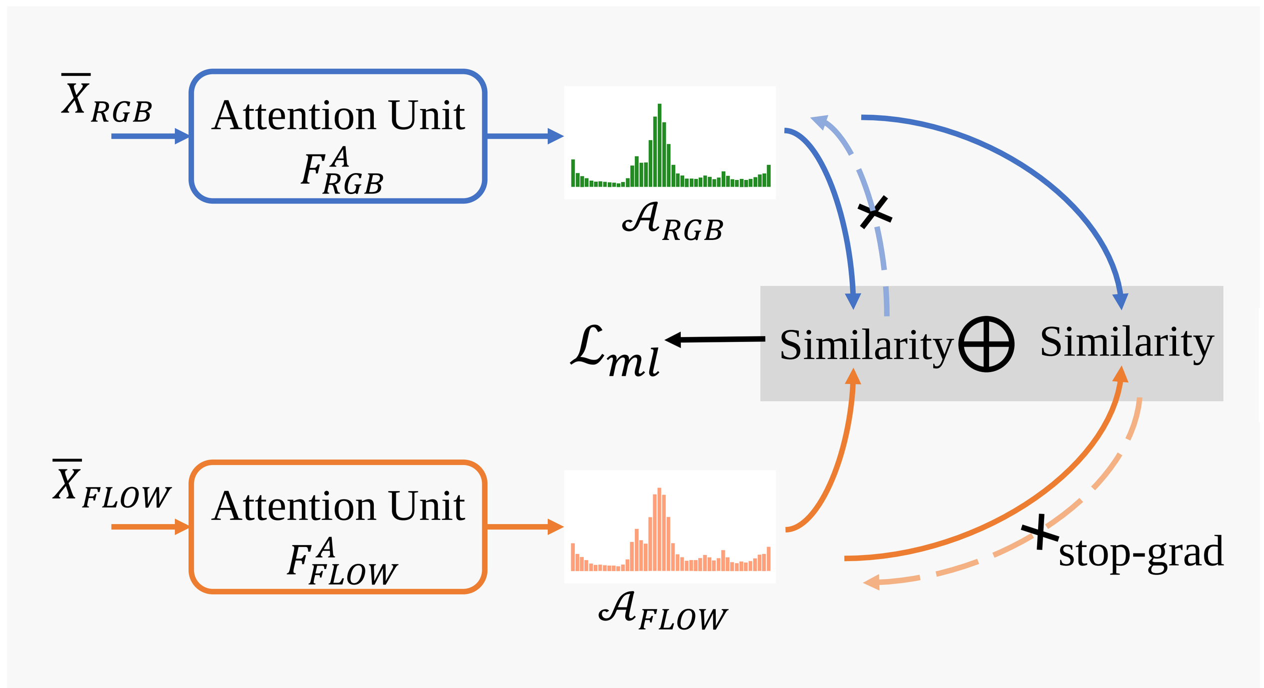
➢ Cross-modal Consensus Module



Supposed RGB is the main modality, while FLOW is the auxiliary one. CCM is to leverage the modal-wise correlation to recalibrate the features.

1. To get the global context from the main modality, a global aggregation function $\Psi(\cdot)$, and a projection function $F^G(\cdot)$ are used to get channel-wise descriptor M^G: $X_g = \psi(X_{RGB})$,
$$M^G = F^G(X_g).$$
2. To get the cross-modal local-specific information, a projection $F^L(\cdot)$ is used.
$$M^L = F^L(X_{FLOW}).$$
3. Then, recalibrate the features by making modal-wise consensus.
$$M = M^G \otimes M^L,$$
$$\overline{X}_{RGB} = \sigma(M) \otimes X_{RGB},$$

➢ Mutual Learning between Dual Modal-specific Attention Units



After feature re-calibration, each modality is equipped with a modal-specific attention unit to generate foreground attention weights $\mathcal{A}_{RGB}$ and $\mathcal{A}_{FLOW}$.

$$\mathcal{A}_{RGB} = F^A_{RGB}(\overline{X}_{RGB}),$$

Considering the modal-wise consistency, $\mathcal{A}_{RGB}$ and $\mathcal{A}_{FLOW}$ can learn from each other via mutual learning.

$$\mathcal{L}_{ml} = \alpha\delta(\mathcal{A}_{RGB}, \phi(\mathcal{A}_{FLOW})) + (1-\alpha)\delta(\mathcal{A}_{FLOW}, \phi(\mathcal{A}_{RGB})), \quad (6)$$

where $\alpha = 0.5$, and $\delta(\cdot)$ is a gradient stopping function.

Then we can get final attention weights as $\mathcal{A}$.

$$\mathcal{A} = \frac{\mathcal{A}_{RGB} + \mathcal{A}_{FLOW}}{2}.$$

➢ Optimizing Process

- Multiple instance learning with / without background suppression. $\mathcal{L}^*_{mil}$ are original Top-K aggregation multiple instance learning loss but applied on different classification scores.

$$\overline{\mathcal{S}} = \mathcal{A} \otimes \mathcal{S}.$$
$$\mathcal{L}_{mil} = \mathcal{L}^{org}_{mil} + \mathcal{L}^{supp}_{mil}$$

- As additional background class in class activation map (CAM) also indicates the foreground probabilities, they can learn from each other, too.

$$\mathcal{L}_{oppo} = \frac{1}{3}(|\mathcal{A}_{RGB} + s_{c+1} - 1| + |\mathcal{A}_{FLOW} + s_{c+1} - 1| + |\mathcal{A} + s_{c+1} - 1|),$$

- To make the attention weights sparse, $\mathcal{L}_{norm}$ is introduced.

$$\mathcal{L}_{norm} = \frac{1}{3}(||\mathcal{A}_{RGB}||_1 + ||\mathcal{A}_{FLOW}||_1 + ||\mathcal{A}||_1),$$

- Besides, $\mathcal{L}_{cas}$ is used to learn more discriminative features via contractive learning. Then we get the overall optimization target $\mathcal{L}$.

$$\mathcal{L} = \mathcal{L}_{mil} + \mathcal{L}_{cas} + \mathcal{L}_{ml} + \lambda_1\mathcal{L}_{oppo} + \lambda_2\mathcal{L}_{norm}$$

Experiments

Supervision	Method	mAP@IoU (%) 0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	AVG mAP (%) 0.1:0.5	0.1:0.7	0.1:0.9
Fully	S-CNN [39] (2016)	47.7	43.5	36.3	28.7	19.0	10.3	5.3	-	-	35.0	24.3	-
	SSN[50] (2017)	60.3	56.2	50.6	40.8	29.1	-	-	-	-	47.4	-	-
	BSN [21] (2018)	-	-	53.5	45.0	36.9	28.4	20.0	-	-	-	-	-
	TAL-Net [3] (2018)	59.8	57.1	53.2	48.5	42.8	**33.8**	**20.8**	-	-	52.3	45.1	-
	P-GCN[47] (2019)	69.5	**67.5**	**63.6**	**57.8**	**49.1**	-	-	-	-	**61.5**	-	-
Weakly†	CMCS[22] (2019)	57.4	50.8	41.2	32.1	23.1	15.0	7.0	-	-	40.9	32.4	-
	STAR [45] (2019)	68.8	60.0	48.7	34.7	23.0	-	-	-	-	47.4	-	-
	3C-Net [28] (2019)	59.1	53.5	44.2	34.1	26.6	-	8.1	-	-	43.5	-	-
	PreTrimNet [49] (2020)	57.5	50.7	41.4	32.1	23.1	14.2	7.7	-	-	41.0	23.7	-
	SF-Net [25] (2020)	71.0	63.4	53.2	40.7	29.3	18.4	9.6	-	-	51.5	40.8	-
Weakly	BaS-Net [19] (2020)	58.2	52.3	44.6	36.0	27.0	18.6	10.4	3.3	0.4	43.6	35.3	27.9
	Gong *et al.* [9] (2020)	-	-	46.9	38.9	30.1	19.8	10.4	-	-	-	-	-
	DML [13] (2020)	62.3	-	46.8	-	29.6	-	9.7	-	-	-	-	-
	A2CL-PT [26] (2020)	61.2	56.1	48.1	39.0	30.1	19.2	10.6	4.8	1.0	46.9	37.8	30.0
	TSCN [48] (2020)	63.4	57.6	47.8	37.7	28.7	19.4	10.2	3.9	0.7	47.0	37.8	29.9
	ACSNet [23] (2021)	-	-	51.4	42.7	32.4	22.0	11.7	-	-	-	-	-
	HAM-Net [12] (2021)	65.9	59.6	52.2	43.1	32.6	21.9	12.5	4.4*	0.7*	50.7	39.8	32.5
	UM [20] (2021)	67.5	61.2	52.3	43.4	33.7	22.9	12.1	3.9*	0.4*	51.6	41.9	33.0
	CO_2-Net	**70.1**	**63.6**	**54.5**	**45.7**	**38.3**	**26.4**	**13.4**	**6.9**	**2.0**	**54.4**	**44.6**	**35.7**

Table 1: Comparisons of CO_2-Net with other methods on the THUMOS14 dataset. AVG is the average mAP under multiple thresholds, namely, 0.1:0.5:0.1, 0.1:0.7:0.1 and 0.1:0.9:0.1; while † means additional information is adopted in this method, such as action frequency or human pose. * indicates that the results are obtained by contacting the corresponding authors via email.

Supervision	Method	mAP@IoU (%) 0.5	0.75	0.95	AVG
Fully	SSN[50] (2017)	41.3	27.0	6.1	26.6
Weakly†	3C-Net [28] (2019)	35.4	22.9	8.5	21.1
	CMCS [22] (2019)	36.8	22.0	5.6	22.4
Weakly	BaSNet [19] (2020)	38.5	24.2	5.6	24.3
	ActionBytes [14] (2020)	39.4	-	-	-
	DGAM [37] (2020)	41.0	23.5	5.3	24.4
	Gong *et al.* [9] (2020)	40.0	25.0	4.6	24.6
	TSCN [48] (2020)	37.6	23.7	5.7	23.6
	RefineLoc [32] (2021)	38.7	22.6	5.5	23.2
	HAM-Net [12] (2021)	41.0	24.8	5.3	25.1
	UM [20] (2021)	41.2	25.6	6.0	25.9
	ACSNet [23] (2021)	40.1	26.1	**6.8**	26.0
	CO_2-Net	**43.3**	**26.3**	5.2	**26.4**

Table 2: Comparison of our algorithm with other methods on the ActivityNet1.2 dataset. AVG means average mAP from IoU 0.5 to 0.95 with 0.05 increment.

method	Add	Concat	SSMA [41]	SE [11]	CCM
Avg mAP	39.9	39.5	38.0	43.0	**44.6**

Table 6: Comparisons with other multi-modal early fusion methods (*i.e.*, addition and concatenation), SSMA [41] and SE-attention [11] in CO_2-Net in term of average mAP under multiple IoU thresholds {0.1:0.7:0.1}.

Exp	$\mathcal{L}_{mil}$	$\mathcal{L}_{oppo}$	$\mathcal{L}_{ml}$	$\mathcal{L}_{cas}$	$\mathcal{L}_{norm}$	Avg mAP (%)
1	✓					38.1
2	✓	✓				40.0
3	✓	✓	✓			41.4
4	✓	✓	✓	✓		42.8
5	✓	✓		✓	✓	42.6
6	✓	✓	✓	✓	✓	44.6

Table 3: Ablation studies of our algorithm in term of average mAP under multiple IoU thresholds as {0.1:0.7:0.1}.