Generalized Uncertainty-Based Evidential Fusion with Hybrid Multi-Head Attention for Weak-Supervised Temporal Action Localization

Yuanpeng He, Lijian Li, Tianxiang Zhan, Wenpin Jiao, Chi-Man Pun

School of Computer Science, Peking University, Beijing, China

Department of Computer and Information Science, University of Macau, Macau, China

Outline

- Introduction
- Motivations
- Contributions
- Methods
- Experiments
- Conclusions

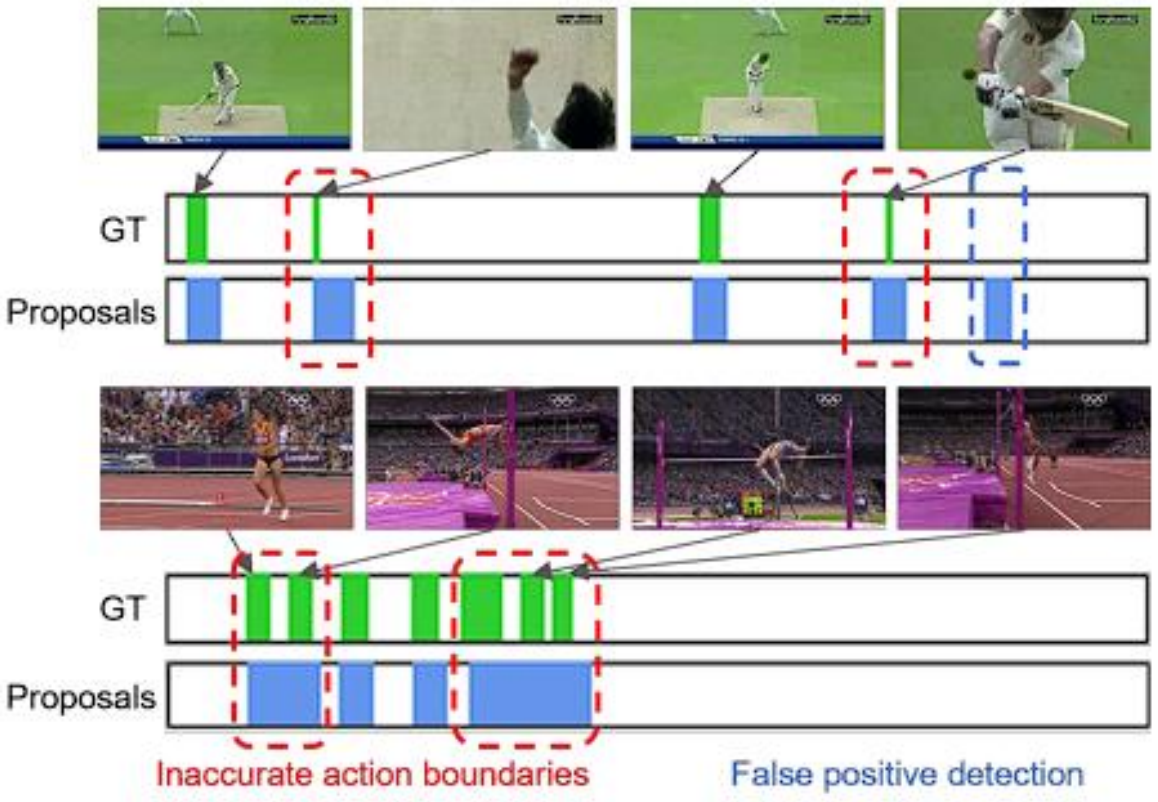
Background

■ Introduction

- Temporal action localization (TAL), which is one of the main tasks of video understanding, aims at localizing the start and end timestamps of action instances in an untrimmed video and classifying them.

- Fully-supervised methods require a huge amount of fine-grained frame-level annotations, which need manual labeling and have annotation bias of annotators.

- To address above issue, weakly-supervised temporal action localization (WTAL) is proposed, which only require easily collected video-level categorical labels.



Background

■ Existing Methods and Motivation

- Four types of existing methods
  - ■ **MIL-based methods**
  - ■ **Attention methods**
  - ■ **Uncertainty-based methods**
  - ■ **Pseudo-label-based methods**

- The shortage of previous researches
  - ■ The primary challenge faced by current WS-TAL methods is action-background ambiguity.
  - ■ Pre-trained model to extract RGB and optical flow features, which encompass a large quantity of redundant information irrelevant to the task.

Contributions

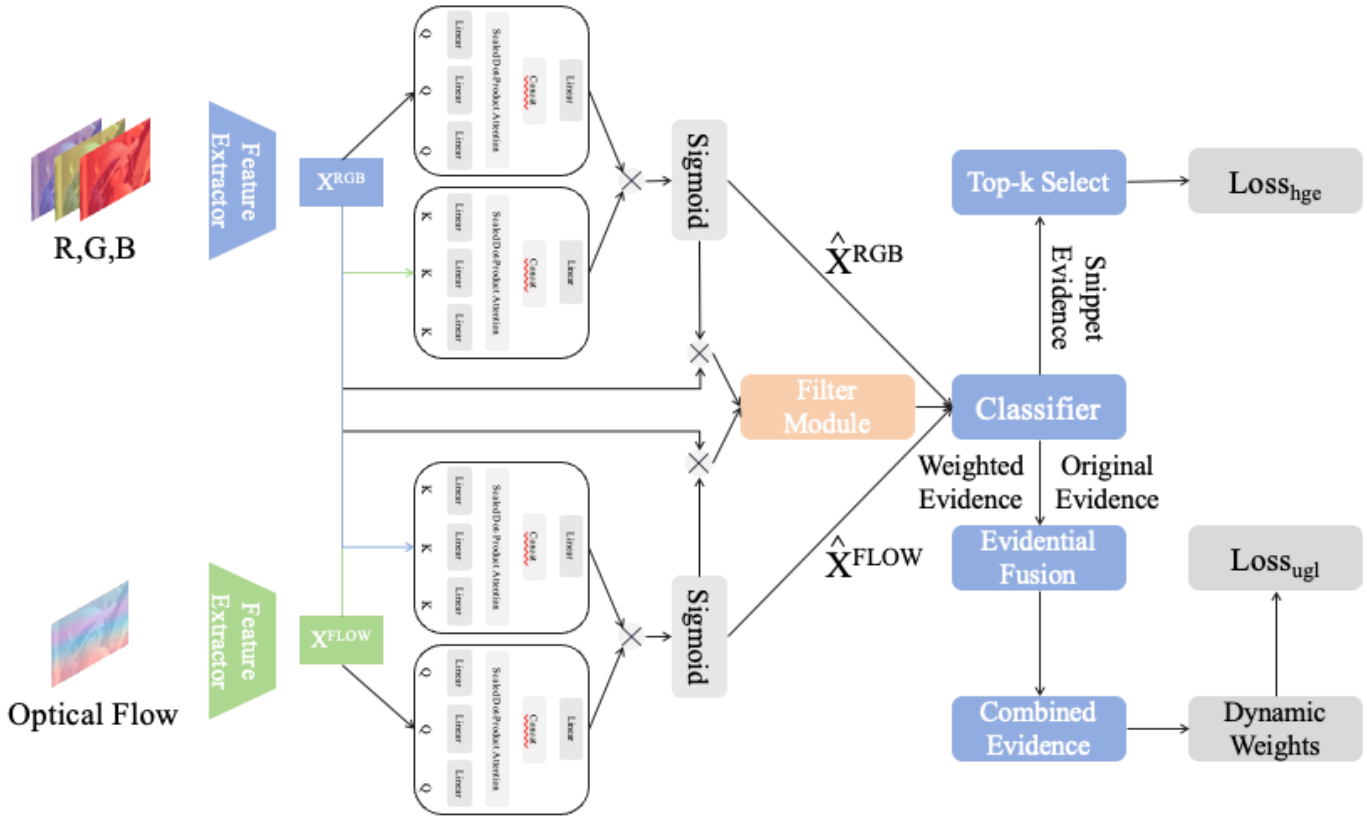
- **We propose a Generalized Uncertainty-Based Evidential Fusion module for the WS-TAL task, which can effectively eliminate action-background ambiguity problem by fusing snippet-level evidences.**

- **A Hybrid Multi-Head Attention module is proposed to enhance the extracted RGB and optical flow features, aligning the feature distribution more appropriately with the requirements of the WS-TAL task.**

- **The results of a large number of experiments conducted on the THUMOS14 dataset demonstrate the excellent performance of our proposed method, surpassing recent state-of-the-art methods.**

Methods

- Two main design of the Methodology
  - Hybrid Multi-Head Attention
  - Deep Learning with Generalized Uncertainty-Based Evidential Fusion



Methods

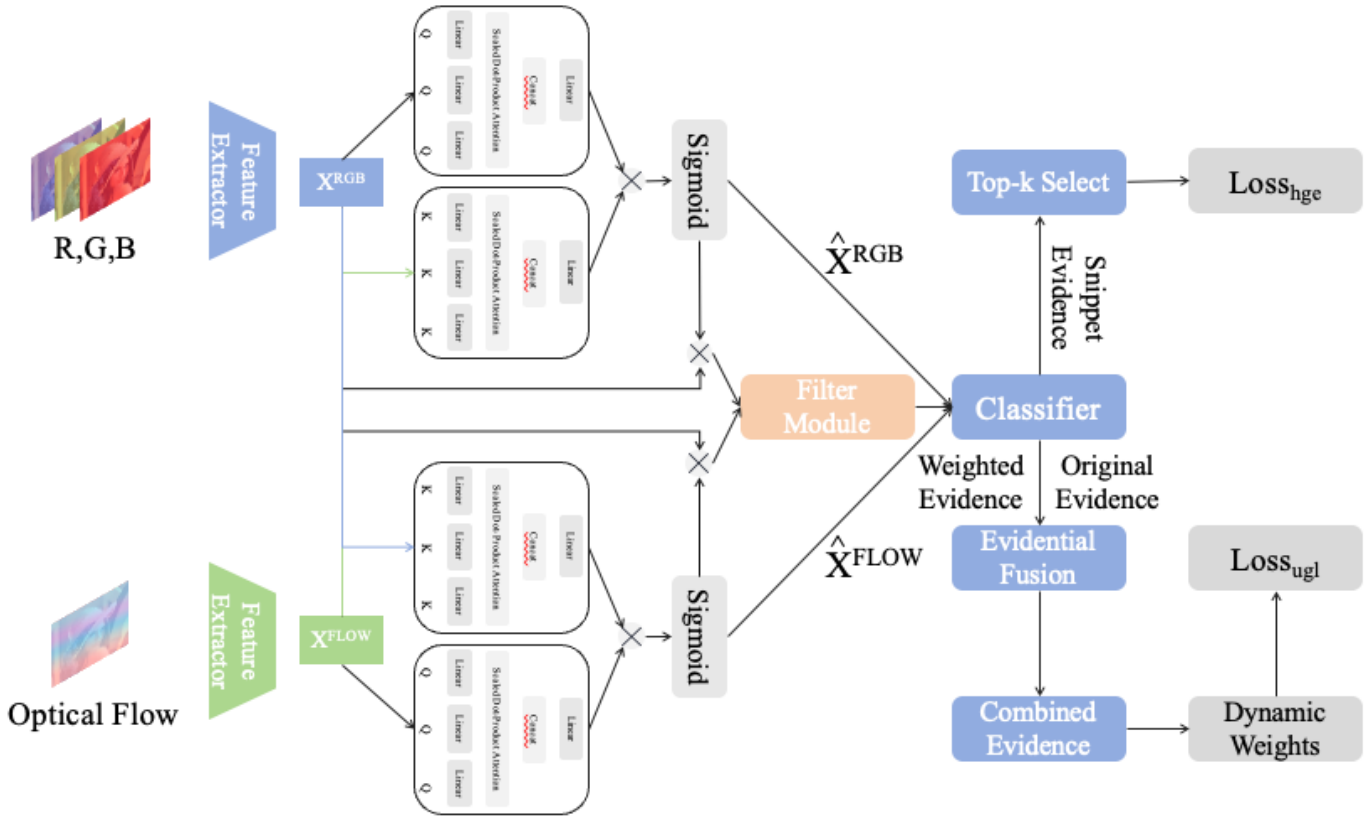
■ Hybrid Multi-Head Attention

Optical flow features and RGB features are fed into two sharing multi-head attention modules MHA to obtain two weights A^{Flow}, $A^{RGB} \in \mathbb{R}^{W}$, which are employed to eliminate task-irrelevant information contained in two initial features

$$A^{Flow}, A^{RGB} = MHA\left(X^{Flow}\right), MHA\left(X^{RGB}\right)$$

$$\hat{X}^{Flow} = X^{Flow} \otimes \sigma\left(A^{Flow} \otimes A^{RGB}\right)$$

$$\hat{X}^{RGB} = X^{RGB} \otimes \sigma\left(A^{RGB} \otimes A^{Flow}\right)$$

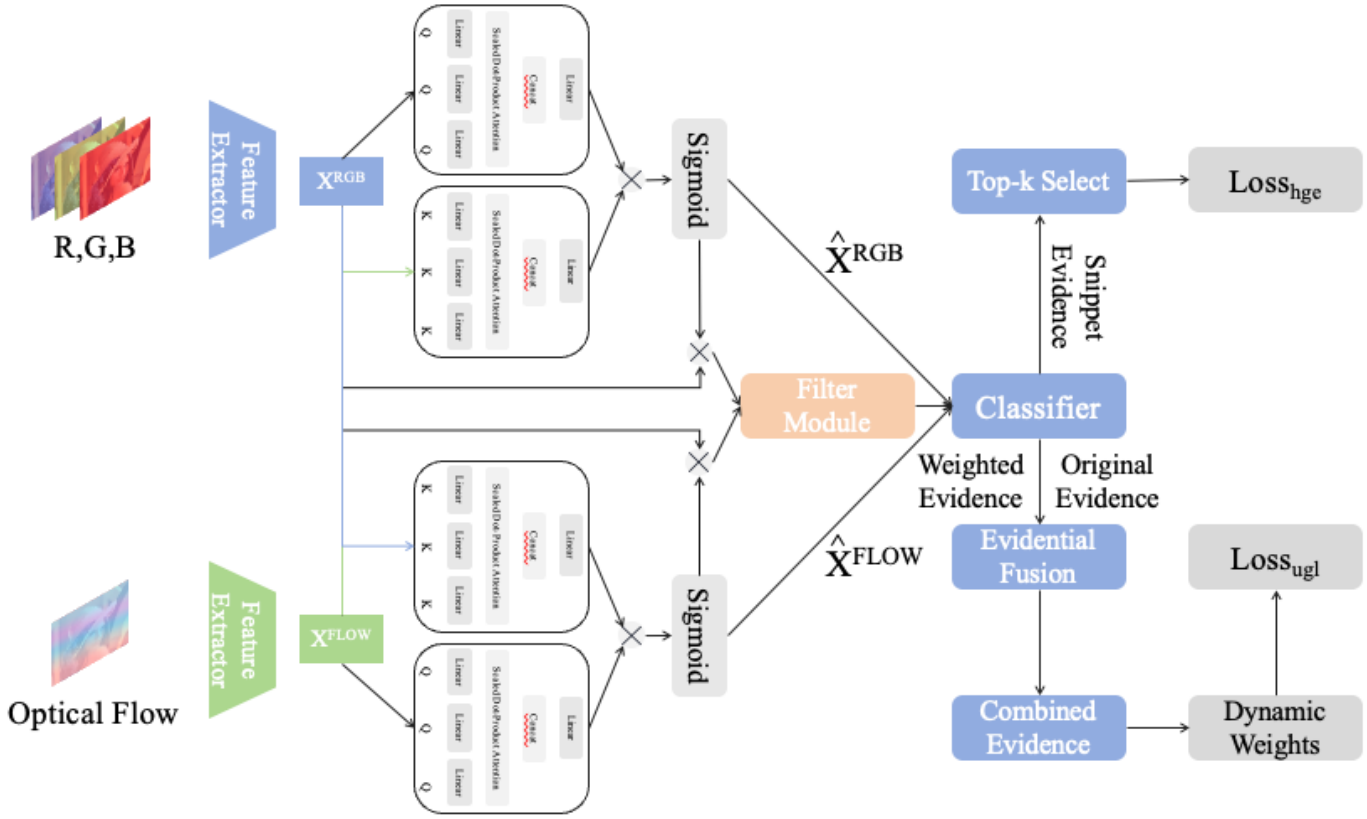


Methods

■ Hybrid Multi-Head Attention

With optimized optical flow $\hat{X}^{Flow}$ and RGB features $\hat{X}^{RGB}$, we intend to use a filtering module f_{attn} which consists of three 1D convolution layers and a sigmoid function, to extract their temporal attention weights.

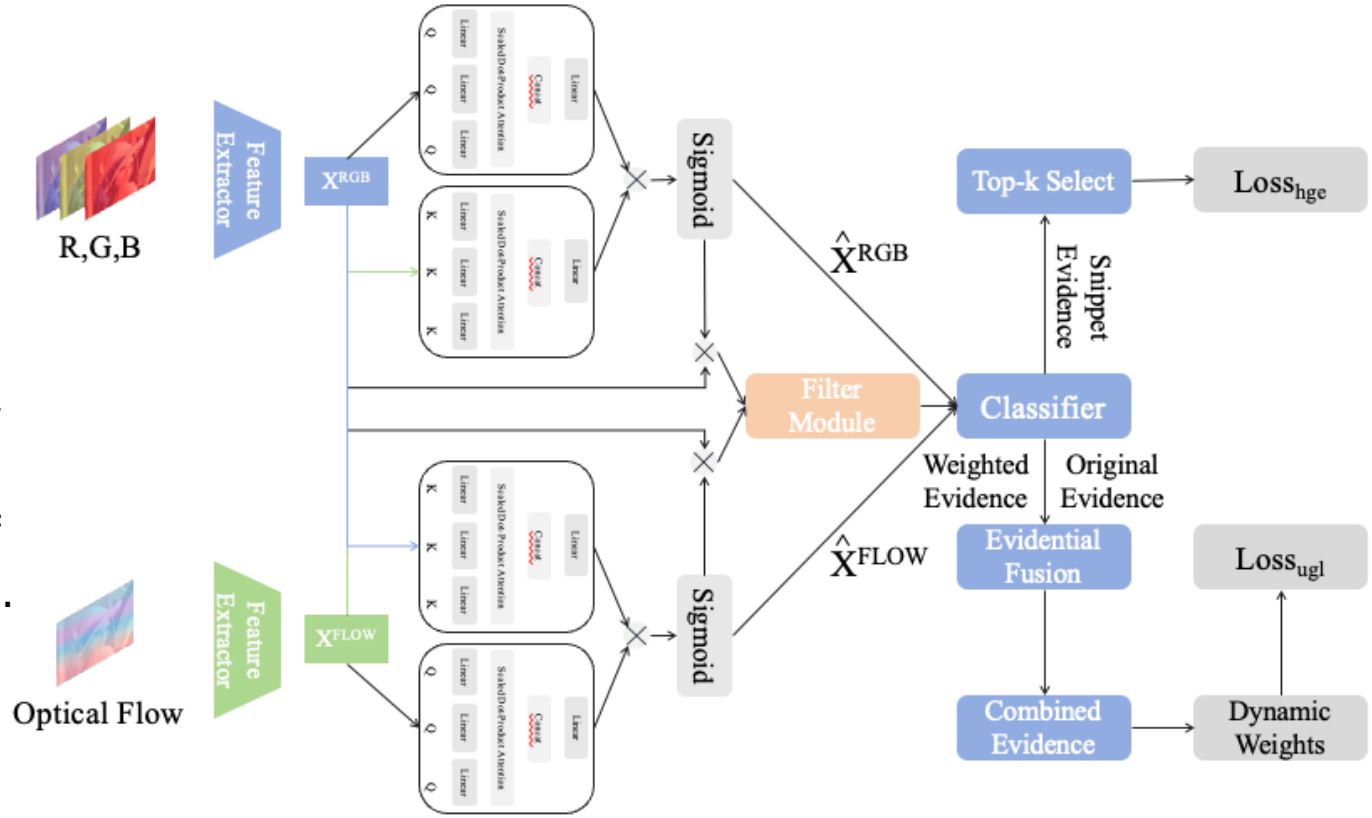
$$\hat{A}^{Flow}, \hat{A}^{RGB} = f_{attn}\left(\hat{X}^{Flow}\right), f_{attn}\left(\hat{X}^{RGB}\right)$$



Methods

Deep Learning with Generalized Uncertainty-Based Evidential Fusion

Evidential deep learning aims to avoid overconfidence of softmax-based classifiers on false predictions. However, according to the original definition of Dempster-Shafer evidence theory, the process of calculation of uncertainty in evidential deep learning is not precise enough to fully represent the level of uncertainty of each complete piece of evidence.



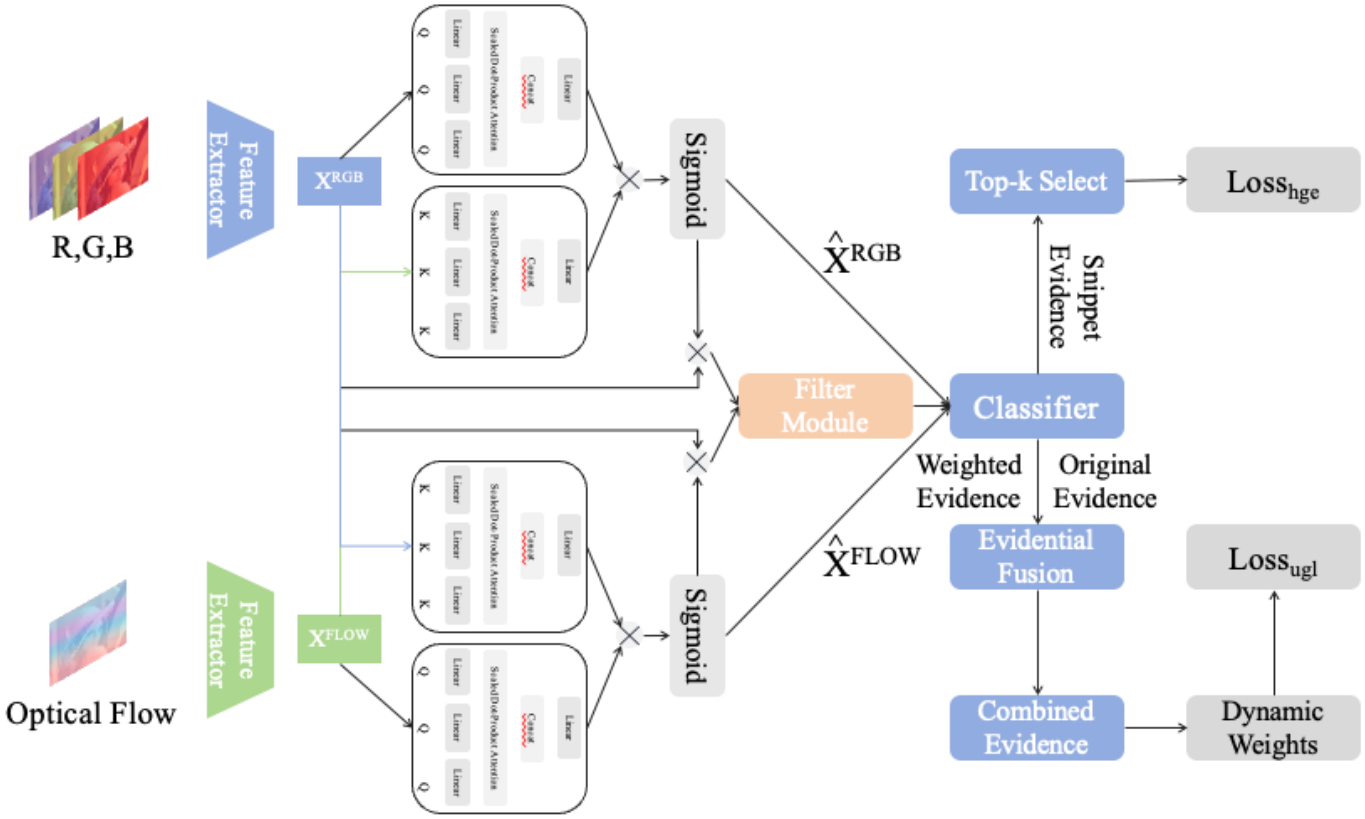
$$m(\{\Theta\}) = U, \Theta = \{p_1, p_2, ., .. p_N\} = p_{N+1}$$

Methods

■ Deep Learning with Generalized Uncertainty-Based Evidential Fusion

In order to better fuse the classification results from different attention weights, we propose a probability distribution fusion strategy at the snippet level.

$$m_{final}(\{p_k\}) = \frac{1}{1-Con}(m_1(\{p_k\}) * m_2(\{p_k\}) + m_1(\{p_k\}) * m_2(\{\Theta\}) + m_1(\{\Theta\}) * m_2(\{p_k\}))$$



Training

■ Loss Function

$$\mathcal{L}_{cla} = Cross\_entropy(\hat{p}, p)$$

$$\mathcal{L}_{\mu ef} = (\Delta \cdot tanh(\sigma(h)\varphi(m_s(\{\Theta)\}) + 1) \sum_{t=1}^{W} L1_{norm}|1 - A_t - z_{t,T+1}|$$

$$\mathcal{L}_{hge} = \sum_{i=1}^{M}(1 - U_{e_s^1}) \sum_{j=1}^{T} \frac{p_j^{(i)}/e_j^{1,(i)}}{\sum_{j=1}^{T} p_j^{(i)}/e_j^{1,(i)}}(logS^{(i)} - log\alpha_j^{(i)})$$

$$\mathcal{L} = \mathcal{L}_{cla} + \lambda_1 \mathcal{L}_{\mu gl} + \lambda_2 \mathcal{L}_{hge}$$

Experiments

We conduct a large amount of experiments to evaluate the proposed method on THUMOS14 dataset. THUMOS14 is composed of 200 validation videos and 213 testing videos with 20 action classes. Besides, The mean Average Precision (mAP) with different Intersection-over-Union (IoU) thresholds, which is regarded as a standard evaluation metric for WS-TAL tasks, is used to evaluate the performance of the proposed model.

Table 1. The experimental results of our model on THUMOS 14 compared with state-of-the-art methods

Supervision	Method	mAP@t-IoU(%) ↑							AVG (0.1-0.5)	AVG (0.3-0.7)	AVG (0.1-0.7)
		0.1	0.2	0.3	0.4	0.5	0.6	0.7			
Fully (-)	SSN[1](ICCV'17)	60.3	60.3	50.6	40.8	29.1	-	-	49.6	-	-
	BSN[2](ECCV'18)	-	-	53.5	45.0	36.9	28.4	**20.0**	-	36.8	-
	GTAN[3](CVPR'19)	69.1	63.7	57.8	47.2	38.8	-	-	55.3	-	-
Weakly (I3D)	FTCL [14](CVPR'22)	69.6	63.4	55.2	45.2	35.6	24.3	12.2	53.8	34.4	43.6
	Huang et al.[4](CVPR'22)	71.3	65.3	55.8	47.5	38.2	25.4	12.5	55.6	35.9	45.1
	ASM-Loc[15](CVPR'22)	71.2	65.5	57.1	46.8	36.6	25.2	13.4	55.4	35.8	45.1
	Li et al.[16](MM'22)	69.7	64.5	58.1	49.9	39.6	27.3	14.2	56.3	37.8	46.1
	DELU[5](ECCV'22)	71.5	66.2	56.5	47.7	40.5	27.2	15.3	56.5	37.4	46.4
	TFE-DCN[6](WACV'23)	72.3	66.5	58.6	49.5	40.7	27.1	13.7	57.5	38.0	46.9
	Wang et al.[17](CVPR'23)	73.0	68.2	60.0	47.9	37.1	24.4	12.7	57.2	36.4	46.2
	Li et al.[7](CVPR'23)	-	-	56.2	47.8	39.3	27.5	15.2	-	37.2	-
	Ren et al.[18](CVPR'23)	71.8	67.5	58.9	49.0	40.0	27.1	15.1	57.4	38.0	47.0
	Ours	**74.5**	**69.1**	**60.3**	**51.2**	**42.1**	**29.4**	15.5	**59.5**	**39.7**	**48.9**

Experiments

Here, we explore the effectiveness of two mentioned modules on the THUMOS14 dataset. Notably, the omission of GUEF results in significant performance degradation, confirming its effectiveness. Besides, only removing HMHA, the performance also shows a moderate degradation within 0.6.

Table 2. The experimental results of ablation study

Exp	*GUEF*	*HMHA*	mAP@IoU(%) ↑				
			0.1	0.3	0.5	0.7	AVG (0.1-0.7)
1	✖	✖	72.1	56.7	38.7	12.3	46.5
2	✖	✔	72.9	58.3	40.2	13.7	47.3
3	✔	✖	73.9	60.0	41.8	14.5	48.4
4	✔	✔	**74.5**	**60.3**	**42.1**	**15.5**	**48.9**

Conclusion

- In this paper, we propose a generalized uncertainty-based evidential fusion and hybrid multi-head attention module, which effectively eliminates action-background ambiguity and filters redundant information from pre-trained features to enable the model to focus on foreground snippets, consequently improving performance.
- Experimental results on the THUMOS14 dataset compared with the latest state-of-the-art methods demonstrate the effectiveness of our proposed method.
- Considering the fact that pseudo-label is also efficacious on WS-TAL tasks, we will conduct further research on this direction.

Thank You!!

Q&A

Our code is available at:

https://github.com/heyuanpengpku/GUEF/tree/main