Journal of Computer Science and Frontier Technologies
ISSN Print: 3104-4204
ISSN Online: 3104-4212
CD ISSN: 3104-4220
https://doi.org/10.63313/JCSFT.9032

FCSC: Weakly-Supervised Temporal Action Localization via Feature Calibration-assisted Sequence Comparison

Lingduo Zhang
Shenyang University of Technology, Shenyang, 110000, China
Email: 15524114460@163.com

How to cite this paper: Zhang, L. D. (2025). FCSC: Weakly-Supervised temporal action localization via feature calibration-assisted sequence comparison. *Journal of Computer Science and Frontier Technologies*, 2(1), 29-41. ISSN Print: 3104-4204; ISSN Online: 3104-4212.
https://doi.org/10.63313/JCSFT.9032
Published: 2025-12-15

Abstract

Weakly-supervised temporal action localization task is to identify action categories and start and end times in unedited videos. How to achieve feature calibration between different modalities in this task, and how to further optimize action boundaries based on the similarity of action common sequences remains an urgent problem to be solved. Based on the above issues, we propose a novel network framework, weakly supervised temporal action localization via feature calibration-assisted sequence comparison (FCSC). The core of the FCSC framework lies in the Multi-Modal Feature Calibration Module (MFCM), which utilizes global and local contextual information from the primary and auxiliary modalities to enhance RGB and FLOW features, respectively, achieving deep feature calibration. In addition, the framework introduces an improved distinguishable edit distance metric to sequence similarity optimize (SSO) and maximum consistent subsequence (MCS) to narrow the gap between classification and localization tasks. After multiple experiments, it has been proven that FCSC achieved maps of 47.7% and 27.9%, respectively on the THUMOS14 and ActiveNet1.2 temporal action recognition benchmark test sets, fully verifying the effectiveness of the model.

Keywords

Weakly - supervised; Multi-modal feature calibration module; Sequence similarity optimize; Maximum consistent subsequence

1. Introduction

Temporal action localization, one of the most fundamental tasks in computer vision, is aimed at identifying action categories and localizing the start and end times of different actions in untrimmed videos [1,2]. In the past few years, action localization models under full supervision have made amazing progress. However, fully supervised methods, based on a large number of frame - level annotations and manual framing, are time - and labor - intensive, which limits their scalability in real life. To

enhance the practicality of action localization models in real life, people have further studied weakly supervised temporal action localization methods that use only video - level labels.

This paper proposes a weakly-supervised temporal action localization framework FCSC with feature enhancement and auxiliary sequence contrast. To address the insufficient semantic information in video features, we design an MFCM. It uses RGB/FLOW features as the main modality to extract global context information, with the other feature as the auxiliary modality to obtain local details. They are deeply fused to generate channel - level descriptors, enhancing semantic richness. An improved discriminative edit distance metric optimizes SSO by calculating the minimum cost of sequence transformation, boosting the stability of action - background separation. Also, we find the MCS of the same action in two untrimmed videos to boost prediction coherence and narrow the classification - localization task gap.

This paper has three main contributions:

- According to the above analysis, we integrate the video natural time series and high - precision temporal resolution, proposing an overall sequence - to - sequence comparison framework. It aims to address the lack of frame - level annotations in weakly - supervised temporal action localization.
- We introduce the multi - modality feature calibration module (MFCM). It fine - tunes the features of each modality, enriches the input features' semantic information, and improves the quality of action localization proposals generated by the model.
- We develop a highly discriminative dynamic programming module, the sequence similarity optimization module (SSO). It enhances the model's precise decoupling ability between actions and backgrounds. Also, the maximum consistent sub - sequence (MCS) improves the model's action recognition accuracy, strengthens the expression of action coherence, and makes the generated video sequences more complete.

2. Related Work

2.1. Weakly-Supervised Action Localization

WS-TAL (Weakly-Supervised Temporal Action Localization) requires only video - level labels for training, attracting much research attention. UntrimmedNet [17]first introduced multiple - instance learning loss, solving the untrimmed video classification problem. Then, STPN [18] added sparsity loss, better distinguishing background segments. It also proposed TCAM for generating action proposals. CO2 [19] designed a cross - modal network for WS - TAL, linking RGB and optical flow features to filter out irrelevant info. But due to the lack of sufficient temporal labels, some studies tried pseudo - labels to guide fully - supervised learning. This often caused many

false action proposals as pseudo - labels are unreliable. To fix this, UGCT [20]introduced uncertainty loss based on pseudo - labels, reducing false samples by filtering out highly uncertain ones. RSKP [21] presented a representative segment summarization and propagation method, generating pseudo - labels only from the most representative segments. ASM - Loc [55]added an uncertainty prediction module, outputting an uncertainty score for each segment to weigh them. Unlike previous methods, this study not only enriches input features via multi - modality feature calibration but also introduces a sequence - contrastive learning mechanism, which greatly enhances the model's ability to recognize and locate actions, achieving better localization results.

2.2. Modal Fusion Enhancement

In recent years, deep neural networks have been widely applied to multi-modal clustering tasks for their strong feature transformation capabilities, with many computer vision models [22,24] integrating multi-modal data to boost performance since different modalities can complement one another. In early research, Ngiam et al.[28]sed deep auto-encoder networks to learn common representations of multi-modal data, improving speech and visual task performance, while subsequent studies [25,23] began combining visual and auditory modalities—videos and audios can mutually enhance each other as their related events often occur simultaneously, as exemplified by Hong et al. [23] who leveraged auditory modality in a multi-head structure to help locate video highlights. Current deep multi-moda subspace clustering algorithms focus on extracting shared features but ignore data distribution within and between modalities, though each modality actually has its own distribution and correlations exist across different modalities. In this study, we enrich input feature semantic information by integrating multi-modal data, combining local and global information, and using feature calibration technology, which enhances the model’s ability to accurately understand action localization.

3. Method

In this study, we propose FCSC framework based WSAL method. As shown in Fig.1, given a pair of video sequences, the goal is to learn an embedding function for each video segment. First, we extract the appearance (RGB) and motion (FLOW) features of each segment via a feature extraction module. Then, the MFCM enhances and fuses these features from different modalities to generate semantically rich and consistent feature representations. These are fed into a MIL module to generate preliminary action proposals. Furthermore, we design complementary learning objectives based on a dynamic programming formula, which includes two key components: SSO and MCS, to capture fine - grained temporal differences. In SSO, gap opening and extension penalties are introduced to help the model identify action boundaries more accurately. MCS uses cosine similarity to locate the longest com-

mon sub - sequence between videos, enabling the transition from classification to localization. These strategies, implemented through dynamic programming, effectively enhance action - background separation and reduce the gap between classification and localization tasks. This provides an accurate and efficient solution for WSAL and significantly improves the precision of video action recognition.

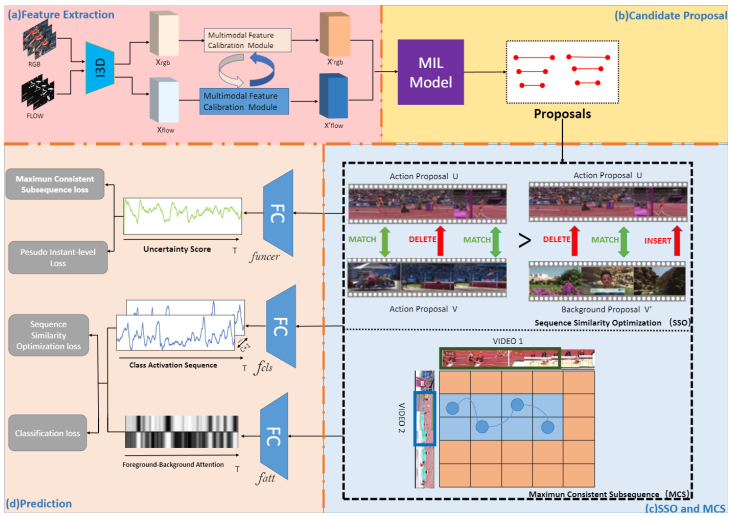

Figure 1. FCFS Network Framework Diagram

3.1. Multi-Modal Feature Calibration Module

In this study, we introduce a multi-modal feature calibration module designed to preprocess information for subsequent network learning tasks and enrich the semantic information of input features. As depicted in Fig.2, this module prioritizes motion modality (FLOW features), with appearance modality (RGB features) serving as an auxiliary component. This design ensures consistent processing even if the roles of the modalities are interchanged, and we primarily focus on FLOW features as the dominant modality.These features are derived from encoders pre-trained on large datasets unrelated to weakly-supervised temporal action localization (WS-TAL), may contain misleading and redundant information. To tackle this challenge, we propose an efficient filtering mechanism between the dominant and auxiliary modalities. This mechanism precisely removes task-irrelevant information from the dominant modality rather than merely concatenating features. Consequently, this approach facilitates a more effective combination of semantic information from both modalities, optimizes feature representation, and enhances model performance in WS-TAL tasks.

We encode modality-specific global context into a video-level feature vector $X_g R^D$.

This vector is derived from the main modality X_{Flow} and aggregated via mean pooling $\psi(\cdot)$ across the temporal dimension. A convolutional layer then captures

channel dependencies to generate a modality-specific global descriptor M^L. The process is as follows:

$$X_g = \psi(X_{RGB}),$$
$$M^L = \mathrm{ConV}(X_g). \quad (1)$$

$$\mathrm{M^G} = \mathrm{Conv}(\mathrm{X_g}) \quad (2)$$

By combining the modality-specific global descriptor MG and the cross-modal local descriptor ML through element-wise multiplication, we obtain the channel descriptor M for feature recalibration. This descriptor M serves as a key tool for optimizing features and enriching semantic information:

$$M = M^G \otimes M^L,$$
$$\mathrm{X}_{FLOW} = \sigma(\mathrm{M}) \otimes \mathrm{X}_{RGB} \quad (3)$$

φ(·) is a Sigmoid function, and "⊗" denotes the element-wise multiplication operator. Note that MG and ML can be used as the ``query'' and ``key'' in an attention mechanism module. The Sigmoid function generates channel-wise recalibration weights to enhance the original main modality features X'Flow.

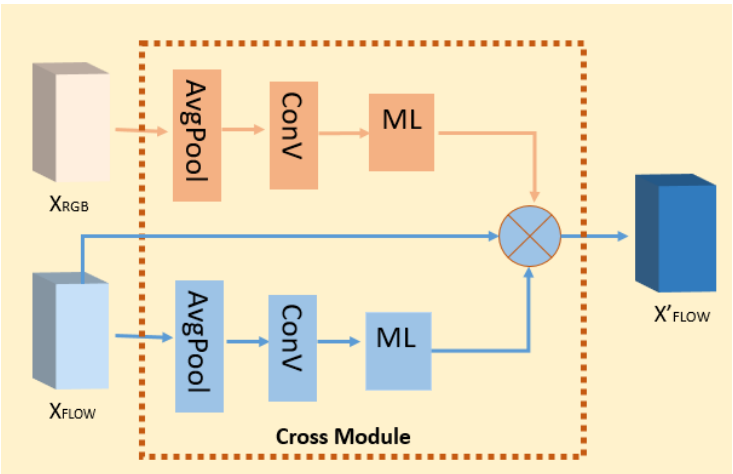

Figure 2. Multi-Modal Feature Calibration Module

3.2. Sequence Similarity Otimize

In this study, we propose a fine-grained temporal sequence contrastive strategy for discriminative action-background separation within the classification localization framework. This strategy shifts from assessing sequence similarity via global feature vector distances to using the minimum cost of transforming one sequence into another. Inspired by the classic edit distance algorithm, we introduce a micro-penalty operation for optimizing sequence similarity, including basic operations like matching, inserting, and deleting, as well as open and extension penalties for accurate sequence transformation cost calculations.

In practical applications, leveraging the learned class activation sequence (CAS) mechanism, the model can precisely generate action proposal set U and background proposal set V. For two proposal sequences of lengths M and N, U = [u1， ...， ui， ...， uM]∈RD×Mand V = [v1， ...， vi， ...， vM]∈RD×N, their similarity is calculated using the following recursive formula:

$$S(i,j) = \mu_{i,j} + max\begin{cases} g_{i,j} + S(i-1,j) - p_{\text{open}} - max(p_{\text{ext}} \cdot e_{i-1,j}, 0) & \text{(Insert)} \\ h_{i,j} + S(i,j-1) - p_{\text{open}} - max(p_{\text{ext}} \cdot e_{i,j-1}, 0) & \text{(Delete)} \end{cases} \quad (4)$$

We define a subsequence similarity score S(i,j) to evaluate the similarity between the i-th position of sequence U and the j-th position of sequence V. We initialize S(0, :) and S(:, 0) to zero for recursive computation. If Ui and Vj match, the similarity score increases; if an insertion or deletion occurs, the score is penalized. Here,popen is the penalty for initial insertions/deletions, and pext penalizes consecutive ones. The indicatorei−1,j is 1 if there are consecutive insertions or matches in rows ei−1,j=1, else 0. We also learn three scalars μi,j, gi,j and hi,j.Taking \(\mu_{i,j}\) and \(g_{i,j}\) as examples, the calculation formulas are as follows:

$$\mu_{i,j} = \sigma_{\mu}\left(\cos\left(\Delta^{\mu}_{i,j}\right)\right), \qquad g_{i,j} = \sigma_{g}\left(\cos\left(\Delta^{g}_{i,j}\right)\right) \quad (5)$$

$\Delta^{\mu}_{i,j}$=[fμ(ui),fμ(vj)], and similarly, $\Delta^{g}_{i,j}$ is defined. $f\mu(\cdot)$, $fg(\cdot)$, and fh(·) are fully connected layers modeling matching, inserting, and deleting operations.σμ and σg extract residuals from their outputs.

$$\mathcal{L}_{\text{SSO}} = \ell(S_{[UV']} - S_{[UV]}) + \ell(S_{[U'V]} - S_{[UV]}) \quad (6)$$

ℓ (x) is the ranking loss, and [UV]denotes same-class action proposals, with U' and V' as background proposals. We use learned importance scores α to select proposals, boosting model performance.

3.3. Maximum Consistent Subsequence

In the previous chapter, we discussed separating actions from backgrounds to boost action classifiers' accuracy. However, WSAL main goal is precise temporal localization of actions using timestamps, which differs from action recognition. To address this, we suggested finding the Maximum Consistent Subsequence (MCS) between two untrimmed videos X and Z. The theory is that if videos have different actions, the MCS is short due to background differences and distinct action categories. Conversely, similar actions lead to a longer MCS, as they consist of comparable temporal segments, ideally aligning with the shorter action instance in the sequences.

Based on these observations, we design a differentiable dynamic programming method to model the Maximum Consistent Subsequence (MCS) between video sequences X and Z . We create a recursive matrix R(T+1)×(T+1), where R(i,j) records

the longest common subsequence length between prefixes Xi and Zj. If Xi and Zj are equal, R(i−1,j−1)+1; otherwise, it retains the previous maximum length. Since segments depicting the same action aren't identical in WSAL tasks, we use a similarity measure between segments to calculate the accumulated sequence length and design a recursive formula for MCS modeling:

$$R(i,j)=\begin{cases}0, & i=0 \text{or} j=0\\ R(i-1,j-1)+w(x_i)\cdot w(z_j)\cdot \cos(xi,zj), & \cos(xi,zj)\geq\tau\\ max\{R(i-1,j),R(i,j-1)\}, & \text{otherwise}\end{cases} \quad (7)$$

In the proposed model, parameter τ is crucial for determining if the i-th segment in video X matches the j-th segment in video Z. Weights w(xi) and w(zj), computed by an attention network, reflect each segment's importance. cos(xi,zj) measures the cosine similarity between segments xi and zj. The model uses these elements to identify the longest common subsequence (MCS) between two video sequences. Via dynamic programming, it calculates r=R(T,T), representing the soft length of the MCS. To optimize LCS learning, a cross-entropy loss function is introduced as a training constraint:

$$\mathcal{L}_{\text{MCS}}=\delta_{xz}\log(r_{[\mathbf{XZ}]})+(1-\delta_{xz})\log(1-r_{[\mathbf{XZ}]}) \quad (8)$$

δxz indicates whether videos X and Z share the same action class.

3.4. Learning and Inference

Training. A multi-stage training method is used to optimize action proposal accuracy. First, train the base model for E epochs to generate initial proposals. Then, conduct E epochs of further training to enhance proposal precision in action localization and duration estimation. This iterative process repeats until proposals stabilize. The loss function includes classification loss, pseudo-supervision loss, two optimization losses, and an instance-aware loss:

$$\mathcal{L}=\mathcal{L}_{cls}+\mathcal{L}_{\text{SSO}}+\mathcal{L}_{\text{MCS}}+\mathcal{L}_{\text{ins}} \quad (9)$$

Inference. Generate action segment proposals, filter out low-confidence ones using a threshold, merge adjacent segments, and apply NMS to remove duplicates.

4. Experiment

Dataset. To validate its effectiveness, the FCSC model was evaluated on two popular action localization benchmark datasets: THUMOS-14 and ActivityNet v1.2. THUMOS-14 contains 20 action classes, with videos ranging from seconds to minutes, and may include multiple action instances. Following previous studies[7,9], 200 validation videos were used for training, and 213 test videos for evaluation. ActivityNet v1.2 covers 200 complex daily activities, with 10,024 training and 4,926 validation videos.

Evaluation Metrics. We used the average precision (mAP) across different temporal

intersection over union (t-IoU) thresholds. For THUMOS14, t-IoU thresholds range from 0.1 to 0.7 in steps of 0.1. For ActivityNet, they range from 0.5 to 0.95 in steps of 0.05.

Experimental Setup.We utilized a pre-trained I3D [8] on Kinetics for feature extraction without fine-tuning.For THUMOS14 and ActivityNet1.2, T was 750 and 75. A pre-trained ACM-Net [7] handled video-level classification. In SSO, CAS-selected clips proposed actions/backgrounds. In MCS, top J CAS-activated clips J=30 for THUMOS14, J=10 for ActivityNet1.2 were used. $f_\mu(\cdot)$and fg(·) had 1024D outputs, with $f_\mu(\cdot)$ matching fg(·) Temperature γ was 10, threshold τ 0.92. The PyTorch 1.9.2 model used Adam optimizer (learning rate 10^{-4}, batch size 16) until loss stabilized.

4.1. Comparison with State-of-the-Art Methods

THUMOS14 Evaluation. As shown in Table.1, FCSC outperforms previous weakly supervised methods on the THUMOS14 dataset, achieving a 47.7\% mAP. It surpasses TSCANet[43]by 2.5\% and even exceeds several fully supervised methods with less annotated data.

ActivityNet1.2 Evaluation. Table.2 shows FCSC also achieves state-of-the-art performance, with a 0.6\% relative improvement over TSCANet[43]. However, the improvement is less significant than on THUMOS14, likely due to ActivityNet's shorter video lengths (1.6 instances per video vs. 15.6 in THUMOS14). Richer temporal information benefits fine-grained temporal sequence comparison, offering greater advantages in complex video scenarios.

Table 1. Comparison of action localization performance between FCSC and state-of-the-art methods on THUMOS14

Supervision	Method	mAP t-IOU(%)							AVG	AVG	AVG
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	(0.1:0.5)	(0.3:0.7)	(0.1:0.7)
	SSN [44]	66.0	59.4	51.9	41.0	29.8	-	-	49.6	-	-
	TAL-Net[45]	59.8	57.1	53.2	48.5	42.8	33.8	20.8	52.3	39.8	45.1
Fully	GTAN[46]	69.1	63.7	57.8	47.2	38.8	-	-	55.3	-	-
	RDFA-S6[47]	-	-	88.7	84.6	78.2	66.6	51.9	-	74.2	-
	ActionMamba[48]	-	-	86.9	83.1	76.9	65.6	50.8	-	72.7	-
	UntrimmedNet[11]	44.4	37.7	28.2	21.1	13.7	-	-	29.0	-	-
	W-TALC[49]	55.2	49.6	40.1	31.1	22.8	-	7.6	39.8	-	-
	3C-Net[37]	59.1	53.5	44.2	34.1	26.6	-	8.1	43.5	-	-
	BaS-Net[50]	58.2	52.3	44.6	36.0	27.0	18.6	10.4	43.6	27.3	35.3
	DGAM[51]	60.0	54.2	46.8	38.2	28.8	19.8	11.4	45.6	29.0	37.0
	CO2-Net[52]	70.1	63.6	54.5	45.7	38.3	26.4	13.4	54.4	35.7	44.6
Weakly	ACSNet [53]	-	-	51.4	42.7	32.4	22.0	11.7	-	-	-
	DCC[54]	69.0	63.8	55.9	45.9	35.7	-	13.7	54.1	-	-
	ASM-Loc [55]	71.2	65.5	57.1	46.8	36.6	-	13.2	55.4	-	-
	ACGNET [56]	68.1	62.6	53.1	44.6	34.7	22.6	12.0	52.6	33.4	42.5
	FTCl[57]	69.6	63.4	55.2	45.2	35.6	23.7	12.2	53.8	34.4	43.6
	RSKP[58]	71.3	65.3	55.8	47.5	38.2	25.4	12.5	55.6	35.9	45.1
	F3-net[59]	69.4	63.6	54.2	46.0	36.5	-	-	53.9	-	-
	ISSF[60]	72.4	66.9	58.4	49.7	41.8	25.5	12.8	57.8	37.6	46.8
	TSCANet [43]	71.7	65.7	56.5	46.4	38.0	25.2	12.5	55.7	35.8	45.2
	MCBA(Ours)	73.2	67.5	59.8	50.6	42.9	26.2	13.3	59.7	38.56	47.7

Table 2. Comparison of action localization performance between FCSC and state-of-the-art methods on ActivityNet 1.2

Method	mAP@t-IOU(%)

	0.5	0.75	0.95	Avg
UntrimmedNet[11]	7.4	3.2	0.7	3.6
W-TALC[49]	37.0	-	-	18.0
CO2-Net[52]	43.3	26.3	5.2	26.4
ACGNET [56]	41.8	26.0	5.9	26.1
ECM[62]	41.0	24.9	6.5	25.5
LPR[65]	43.6	26.0	5.2	26.5
TSCANet[43]	44.8	27.1	6.9	27.3
MCBA(Ours)	46.2	28.3	9.2	27.9

4.2. Ablation Experiments

To prove the superior performance of the FCSC algorithm, we conducted ablation studies on the THUMOS14 dataset.

Ablation Study on Multi-Modal Feature Calibration Module To validate the MFCM's effectiveness, we conducted ablation experiments on the THUMOS14 dataset. Removing MFCM caused significant mAP drops across thresholds (Table 3). This shows MFCM is essential for integrating multi-modal features and boosting model performance.

Ablation Study on Sequence Similarity Optimization The ablation study on SSO confirms its critical role in foreground-background separation. Removing SSO led to significant performance drops (Table 3), highlighting its importance for handling complex background relations and maintaining classifier effectiveness

Ablation Study on Maximum Consistent Subsequence To validate the MCS comparison's efficacy, an ablation study was conducted by removing the MCS module. The significant performance drop, as detailed in Table 3, affirms its critical role in bolstering the model's robust classification performance.

Impact of Multi-step Training Refinement Table 4 demonstrates the impact of proposal refinement steps on model performance. Increased steps lead to notable improvements, highlighting the benefits of iterative refinement for target localization. We adopted three steps as the standard due to performance plateauing beyond this point.

Table 3. Contributions of Each Component

	MCBA		AVG
MFCM	SSO	MCS	0.1:0.7
√			44.2
	√		44.3
		√	43.8
√	√		46.5
√		√	46.8
	√	√	47.2
√	√	√	47.7

Table 4. Multi-step Refinement’s Impact on Model Performance

NUM	mAP@t-IOU(%)			
	0.3	0.5	0.7	AVG
0	54.8	39.5	10.2	37.6

1	57.5	40.3	11.8	44.3
2	58.4	41.2	12.7	45.2
3	59.8	42.9	13.3	47.7
4	59.8	42.9	13.3	47.7

4.3. Figures and Tables

The visualizations below demonstrate the FCSC model's effectiveness in addressing common issues with multi-instance methods. It significantly improves detection of short actions and incomplete localization of "basketball" (Fig.3) and over-localization of "long jump" (Fig.4) through action-aware segment modeling.

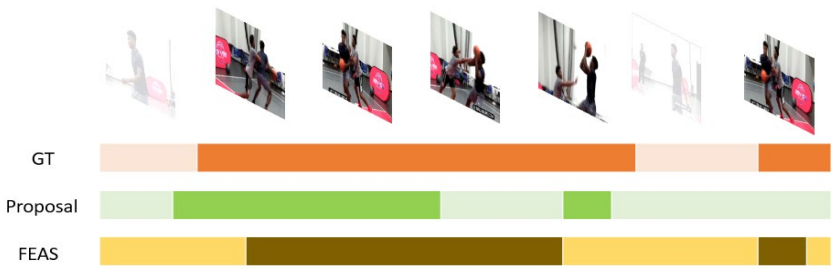

Figure 3. Action Localization Example - Basketball

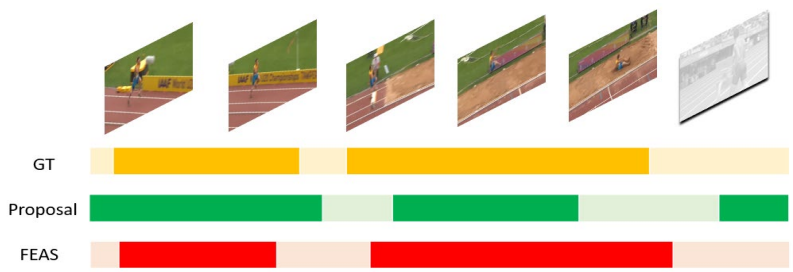

Figure 4. Action Localization Example - Longjump

5. Conclusion

This paper presents a weakly supervised action localization (WSAL) framework utilizing feature calibration and sequence comparison. The Multi-Modal Feature Calibration Module (MFCM) enhances feature integration, while Sequence Similarity Optimization (SSO) and Maximum Consistent Subsequence (MCS) improve action discrimination and continuity. Experiments confirm the modules' effectiveness.

References

[1] Fuchen Long, Ting Yao, Zhaofan Qiu, Xinmei Tian, Jiebo Luo, and Tao Mei. Gaussian temporal awareness networks for action localization. In CVPR, 2019.

[2] Deepak Sridhar, Niamul Quader, Srikanth Muralidharan, Yaoxin Li, Peng Dai, and Juwei Lu. Class semantics-based attention for action detection. In ICCV, 2021.

[3] Mengmeng Xu, Chen Zhao, David S Rojas, Ali Thabet, and Bernard Ghanem. G-tad:

Sub-graph localization for temporal action detection. In CVPR, 2020.

[4] Zixin Zhu, Wei Tang, Le Wang, Nanning Zheng, and Gang Hua. Enriching local and global contexts for temporal action localization. In ICCV, 2021.

[5] Fa-Ting Hong, Jia-Chang Feng, Dan Xu, Ying Shan, and Wei-Shi Zheng. Cross-modal consensus network for weakly supervised temporal action localization. In ACM MM, 2021.

[6] Linjiang Huang, Liang Wang, and Hongsheng Li. Foreground-action consistency network for weakly supervised temporal action localization. In ICCV, 2021.

[7] Zhijun Li Lijun Zhang Fan Lu Alois Knoll Sanqing Qu, Guang Chen. Acm-net: Action context modeling network for weakly-supervised temporal action localization. arXiv:2104.02967, 2021.

[8] Wenfei Yang, Tianzhu Zhang, Xiaoyuan Yu, Tian Qi, Yongdong Zhang, and Feng Wu. Uncertainty guided collaborative training for weakly supervised temporal action detection. In CVPR, 2021.

[9] Sujoy Paul, Sourya Roy, and Amit K Roy-Chowdhury. W- talc: Weakly-supervised temporal activity localization andclassification. In ECCV, 2018.

[10] Baifeng Shi, Qi Dai, Yadong Mu, and Jingdong Wang. Weakly-supervised action localization by generative atten- tion modeling. In CVPR, 2020.

[11] Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. Untrimmednets for weakly supervised action recognition and detection. In CVPR, 2017.

[12] Yunlu Xu, Chengwei Zhang, Zhanzhan Cheng, Jianwen Xie, Yi Niu, Shiliang Pu, and Fei Wu. Segregated temporal as- sembly recurrent networks for weakly supervised multiple action detection. In AAAI, 2019.

[13] Can Zhang, Meng Cao, Dongming Yang, Jie Chen, and Yuexian Zou. Cola: Weakly-supervised temporal action localization with snippet contrastive learning. In CVPR, 2021.

[14] Junwei Ma, Satya Krishna Gorti, Maksims Volkovs, and Guangwei Yu. Weakly supervised action selection learning in video. In CVPR, 2021

[15] Wang Luo, Tianzhu Zhang, Wenfei Yang, Jingen Liu, Tao Mei, Feng Wu, and Yongdong Zhang. Action unit memory network for weakly supervised temporal action localization. In CVPR, 2021.

[16] Yuan Liu, Jingyuan Chen, Zhenfang Chen, Bing Deng, Jian- qiang Huang, and Hanwang Zhang. The blessings of unla- beled background in untrimmed videos. In CVPR, 2021.

[17] Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. Untrimmednets for weakly supervised action recognition and detection. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pages 4325– 4334, 2017.

[18] Phuc Nguyen, Ting Liu, Gautam Prasad, and Bohyung Han. Weakly supervised action localization by sparse temporal pooling network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6752– 6761, 2018.

[19] Fa-Ting Hong, Jia-Chang Feng, Dan Xu, Ying Shan, and Wei-Shi Zheng. Cross-modal consensus network for weakly supervised temporal action localization. In Proceedings of the 29th ACM International Conference on Multimedia, pages 1591–1599, 2021.

[20] Wenfei Yang, Tianzhu Zhang, Xiaoyuan Yu, Tian Qi, Yongdong Zhang, and Feng Wu. Uncertainty guided collaborative training for weakly supervised temporal action detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 53–63, 2021.

[21] Linjiang Huang, Liang Wang, and Hongsheng Li. Weakly supervised temporal action localization via representative snippet knowledge propagation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3272–3281, 2022.

[22] Cheng Deng, Zhaojia Chen, Xianglong Liu, Xinbo Gao, and Dacheng Tao. 2018. Triplet-based deep hashing network for cross-modal retrieval. TIP (2018).

[23] Fa-Ting Hong, Xuanteng Huang, Wei-Hong Li, and Wei-Shi Zheng. 2020. MININet: Multi-

ple Instance Ranking Network for Video Highlight Detection. In ECCV.

[24] Ya Jing, Wei Wang, Liang Wang, and Tieniu Tan. 2020. Cross-Modal Cross-Domain Moment Alignment Network for Person Search. In CVPR.

[25] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. 2020. Self-supervised learning of audio-visual objects from video. arXiv (2020)

[26] Anyi Rao, Linning Xu, Yu Xiong, Guodong Xu, Qingqiu Huang, Bolei Zhou, and Dahua Lin. 2020. A Local-to-Global Approach to Multi-Modal Movie Scene Segmentation. In CVPR.

[27] Jonathan Munro and Dima Damen. 2020. Multi-Modal Domain Adaptation for Fine-Grained Action Recognition. In CVPR.

[28] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. 2011. Multimodal deep learning. In ICML.

[29] Kaidi Cao, Jingwei Ji, Zhangjie Cao, Chien-Yi Chang, and Juan Carlos Niebles. Few-shot video classification via temporal alignment. In CVPR, 2020.

[30] Chien-Yi Chang, De-An Huang, Yanan Sui, Li Fei-Fei, and Juan Carlos Niebles. D3tw: Discriminative differentiable dynamic time warping for weakly supervised action alignment and segmentation. In CVPR, 2019.

[31] Xiaobin Chang, Frederick Tung, and Greg Mori. Learning discriminative prototypes with dynamic time warping. In CVPR, 2021.

[32] Isma Hadji, Konstantinos G Derpanis, and Allan D Jepson. Representation learning via global temporal alignment and cycle-consistency. In CVPR, 2021.

[33] Xingyu Cai and Tingyang Xu. Dtwnet: a dynamic timewarping network. NeurIPS, 2019.

[34] Marco Cuturi and Mathieu Blondel. Soft-dtw: a differentiable loss function for time-series. In ICML, 2017.

[35] Nikita Dvornik, Isma Hadji, Konstantinos G Derpanis, Animesh Garg, and Allan D Jepson. Drop-dtw: Aligning common signal between sequences while dropping outliers. arXiv:2108.11996, 2021.

[36] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Temporal cycleconsistency learning. In CVPR, 2019.

[37] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In CVPR.

[38] Ziyi Liu, Le Wang, Wei Tang, Junsong Yuan, Nanning Zheng, and Gang Hua. Weakly supervised temporal action localization through learning explicit subspaces for action and context. In AAAI, 2021.

[39] Haiping Zhang, Haixiang Lin, Dongjing Wang, Dongyang Xu, Fuxing Zhou, Liming Guan, Dongjing Yu, Xujian Fan. TSCANet: a two stream context aggregation network for weakly supervised temporal action localization.

[40] Zhai Y, Wang L, Tang W, Zhang Q, Yuan J, Hua G (2020) Two-stream Consensus Network for Weakly-supervised Temporal Action Localization. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16, pp. 37–54. Springer

[41] Chao Y-W, Vijayanarasimhan S, Seybold B, Ross DA, Deng J, Sukthankar R (2018) Rethinking the Faster R-cnn Architecture for Temporal Action Localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1130–1139

[42] Long F, Yao T, Qiu Z, Tian X, Luo J, Mei T (2019) Gaussian Temporal Awareness Networks for Action Localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 344–353

[43] Lee S, Jung J, Oh C, Yun S (2024) Enhancing Temporal Action Localization: Advanced s6 Modeling with Recurrent Mechanism.

[44] Chen G, Huang Y, Xu J, Pei B, Chen Z, Li Z, Wang J, Li K, Lu T, Wang L (2024) Video Mamba Suite: State Space Model as a Versatile Alternative for Video Understanding.

[45] Paul S, Roy S, Roy-Chowdhury AK (2018) W-TALC: Weakly-Supervised Temporal Activity Localization and Classifcation, pp. 588–607

[46] Nguyen P, Liu T, Prasad G, Han B (2018) Weakly Supervised Action Localization by Sparse Temporal Pooling Network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6752–6761

[47] Shi B, Dai Q, Mu Y, Wang J (2020) Weakly-supervised Action Localization by Generative Attention Modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1009–1019

[48] Hong F-T, Feng J-C, Xu D, Shan Y, Zheng W-S (2021) Cross-modal Consensus Network for Weakly Supervised Temporal Action Localization

[49] Liu Z, Wang L, Zhang Q, Tang W, Yuan J, Zheng N, Hua G (2021) Acsnet: Action-context Separation Network for Weakly Supervised Temporal Action Localization.