Temporal Action Localization in the Deep Learning Era: A Survey

Binglu Wang *Member, IEEE,* Yongqiang Zhao *Member, IEEE,* Le Yang, Teng Long *Fellow, IEEE,* and Xuelong Li *Fellow, IEEE*

Abstract—The temporal action localization research aims to discover action instances from untrimmed videos, representing a fundamental step in the field of intelligent video understanding. With the advent of deep learning, backbone networks have been instrumental in providing representative spatiotemporal features, while the end-to-end learning paradigm has enabled the development of high-quality models through data-driven training. Both supervised and weakly supervised learning approaches have contributed to the rapid progress of temporal action localization, resulting in a multitude of methods and a large body of literature, making a comprehensive survey a pressing necessity. This paper presents a thorough analysis of existing action localization works, offering a well-organized taxonomy that highlights the strengths and weaknesses of each strategy. In the realm of supervised learning, in addition to the anchor mechanism, we introduce a novel classification mechanism to categorize and summarize existing works. Similarly, for weakly supervised learning, we extend the traditional pre-classification and post-classification mechanisms by providing a fresh perspective on enhancement strategies. Furthermore, we shed light on the bottleneck of confidence estimation, a critical yet overlooked aspect of current works. By conducting detailed analyses, this survey serves as a valuable resource for researchers, providing beneficial guidance to newcomers and inspiring seasoned researchers alike.

Index Terms—Temporal action localization, supervised learning, weakly supervised learning, deep learning, survey

1 Introduction

In the modern era, video has emerged as a versatile, cost-effective, and powerful medium for transporting information. Various scenarios, such as traffic monitoring, sports competitions, and film production, continuously generate a vast number of videos. While these videos contain informative action segments, they are often interspersed with lengthy sequences of irrelevant backgrounds. Such videos are referred to as untrimmed videos [1], [2], in contrast to human-trimmed videos that solely focus on informative action segments. The objective of temporal action localization is to address this challenge by efficiently extracting meaningful action instances and providing their respective starting time, ending time, and classification label. The task of temporal action localization holds fundamental importance in intelligent video analysis and offers significant contributions to a multitude of applications. Notable applications include video editing [3], video content analysis [4], highlight extraction [5], video summarization [6], [7], video-based recommendation [8], industrial video analysis [9], abnormal behavior detection [10], smart surveillance [11], and human-robot interaction [12]. Through robust temporal action localization, these applications stand to benefit from improved efficiency, accuracy, and automation.

To effectively detect action instances in untrimmed videos, temporal action localization algorithms necessitate robust spatio-temporal modeling. This entails the simultaneous consideration of appearance cues within individual frames and the temporal evolution across multiple neighboring frames. Fig. 1 illustrates how the community approaches temporal action localization with distinct supervision levels, namely, fully supervised and weakly supervised methodologies.

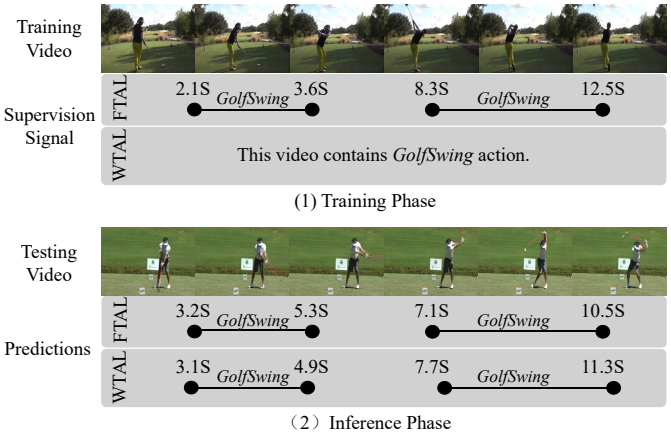

Fig. 1: Illustration of the supervised temporal action localization task (TAL) and the weakly supervised temporal action localization task (WTAL), where the difference lies in the supervision signal.

This survey is dedicated to exploring the advancements in temporal action localization within the context of deep learning. Over the past seven years, the research landscape for temporal action localization has been remarkably dynamic, with a substantial volume of literature emerging since its inception [1]. As depicted in Fig. 2, the supervised methods have undergone impressive performance improvements. Notably, under an Intersection over Union (IoU) threshold of 0.5, the supervised approach has seen its performance surge from 19.0% [1] to an impressive 72.9% [24], representing a remarkable fourfold increase over the past seven-

Binglu Wang and Teng Long are with School of Information and Electronics, Beijing Institute of Technology, Beijing, P.R. China. Binglu Wang is also with Northwestern Polytechnical University, Xi'an 710072, P.R. China. (email: wbl921129@gmail.com).
Yongqiang Zhao, Le Yang and Xuelong Li are with Northwestern Polytechnical University, Xi'an 710072, P.R. China. Corresponding author: Le Yang (email: nwpuyangle@gmail.com) and Xuelong Li (email: li@nwpu.edu.cn).
This work was supported in part by the China Postdoctoral Science Foundation under Grant 2022M710393, 2023M733387, 2022TQ0035, 2023TQ0344.

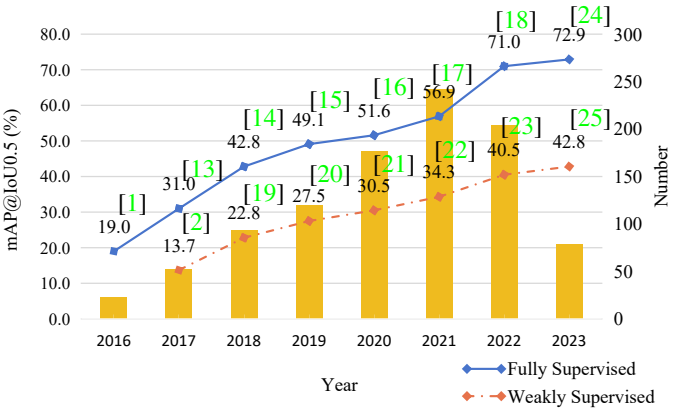

Fig. 2: Performance development for both supervised methods and weakly supervised methods, measured by mAP under IoU threshold 0.5 on THUMOS14 dataset.

year period. While the task of weakly supervised temporal action localization was proposed subsequently, its reliance on video-level labels presents unique challenges. Nonetheless, significant progress has been achieved in this area as well. Commencing with an initial performance of 13.7% [2], the weakly supervised approach has now reached a commendable 42.8% [25].

In times of rapid change, it becomes imperative to summarize and assess existing research to understand the reasons behind advancements, gather valuable insights, identify current methodological bottlenecks, and outline future research directions. Temporal action localization plays a crucial role in the domain of video understanding, alongside foundational techniques like action recognition. Despite action recognition being extensively covered in multiple high-quality surveys [29], [30], [31], the temporal action localization domain lacks a comprehensive survey. As shown in Table 1, although Xia et al. [26] recognize the need to review existing works, their study focus on methods preceding 2018, resulting in the omission of recent progress and insufficient in-depth analyses. Recently, Elahe et al. [27] provide a summary of both fully-supervised and weakly-supervised methods for temporal action localization, including discussions on related settings such as spatio-temporal action detection. However, Elahe et al. [27] primarily elaborate on well-known paradigms, such as categorizing methods into anchor-based and anchor-free pipelines, without introducing a brand-new categorization pipeline. In contrast, our work offers multiple categorizations and highlights a key bottleneck present in existing action localization algorithms.

To address this gap comprehensively, we present a thorough survey of representative action localization works in the deep learning era. Our survey encompasses multiple crucial aspects to provide a comprehensive analysis of the field. Firstly, we systematically categorize existing works into distinct groups, developing a well-defined taxonomy that offers a clear and structured overview of the various paradigms employed in temporal action localization. Alongside this categorization, we conduct a meticulous analysis of the strengths and weaknesses associated with each paradigm, providing valuable insights into their performance. Secondly, throughout the survey, we emphasize the characteristics of influential works that have significantly contributed to the advancement of temporal action localization research. By recognizing these pioneering contributions, we shed light on the prominent trends and developments within the field. Furthermore, our investigation identifies a critical performance bottleneck in current temporal action localization methods, primarily stemming from challenges in accurately estimating confidence scores. Finally, drawing from our comprehensive analysis, we conclude by suggesting promising research directions for subsequent investigations in temporal action localization. These insights are intended to inspire and guide future research endeavors, with the ultimate goal of driving continued advancements in the field.

- An in-depth review for the supervised temporal action localization task from multiple perspectives. Besides the classical anchor perspective, we propose a new classification perspective, and categorize existing works into the frame classification pipeline or the proposal classification pipeline.
- An in-depth review for the weakly supervised temporal action localization task. Apart from the traditional perspective, we review existing works from the enhancement perspective, and categorize them into enhanced classifier, enhanced feature, and enhanced attention pipeline.
- We identify a potential bottleneck in the process of confidence score estimation within the context of temporal action localization. This identification sheds light on introducing refined algorithms for evaluating prediction quality, ultimately elevating the performance of action localization to a higher level.

Our survey endeavors to offer a comprehensive and insightful analysis of the temporal action localization field by thoroughly exploring the strengths, weaknesses, and research trends of various methodologies. The subsequent sections of the paper are structured as follows. Section 2 serves as an introduction to the background of temporal action localization, providing essential context to the readers. In Section 3, we delve into the supervised temporal action localization task, tracing its evolution and presenting multiple categorizations. This section offers a comprehensive discussion of the pros and cons associated with different approaches in this domain. Similarly, Section 4 is dedicated to the discussion of the weakly supervised temporal action localization task. We explore its evolution and various categorizations, engaging in an in-depth analysis of the advantages and limitations of existing methodologies. Next, in Section 5, we identify a critical bottleneck within current methods, specifically the challenge of accurately estimating confidence scores, and discuss other promising research directions. Finally, in Section 6, we draw meaningful conclusions from our analysis, summarizing the key findings and highlighting potential avenues for further investigation. Through this well-structured approach, our survey aims to contribute to the advancement of temporal action localization research while providing valuable insights to the scholarly community.

2 Background

In this section, we provide a comprehensive research background for the temporal action localization task. We commence by presenting a general formulation of the task, capturing its essential characteristics and objectives. Subsequently, we explore and differentiate between two distinct settings: the supervised setting and the weakly supervised setting, each offering unique challenges and approaches for temporal action localization. Additionally, we offer a concise overview of the historical development of temporal action localization. By examining its evolutionary journey, we gain valuable insights into the progression of methods and key milestones that have shaped the field over time. Through this detailed exposition, our aim is to establish a solid foundation of knowledge that sets the stage for the subsequent discussions and analyses in this survey.

TABLE 1: Summary of previous temporal action localization surveys about the temporal action localization task.

Title	Year	Venue	Description
A Survey on Temporal Action Localization [26]	2020	IEEE Access	This paper reviews techniques and models for temporal action localization up to 2019.
Deep Learning-based Action Detection in Untrimmed Videos: A Survey [27]	2022	TPAMI	This paper reviews works for temporal action detection and spatio-temporal action detection.
Weakly-supervised Temporal Action Localization: A Survey [28]	2022	NCA	This paper reviews temporal action localization under only video-level labels .

2.1 Notation and taxonomy

In this context, we are given a dataset $\mathcal{X}$ comprising N videos, denoted as $\mathcal{X} = \{\mathbf{X}_i\}_{i=1}^{N}$. Each video $\mathbf{X}_i$ is accompanied by a corresponding supervision label $\mathbf{Y}_i$. During the training phase, the algorithm processes video $\mathbf{X}_i$ and generates predictions, which are then compared with the respective supervision label $\mathbf{Y}_i$ to guide the learning process.

$$l = \mathcal{L}(\mathbf{\Phi}(\mathbf{X}_i|\mathbf{W}), \mathbf{Y}_i), \tag{1}$$

The loss function $\mathcal{L}(\cdot)$ is utilized to quantify the discrepancy between the predictions made by the algorithm, where $\mathbf{\Phi}(\mathbf{X}_i|\mathbf{W})$ represents the algorithm's predictions, and $\mathbf{W}$ denotes the algorithm's parameter.

The training phase ensures that the algorithm learns to recognize accurate action patterns for each category, leveraging the provided supervision labels. During the inference phase, the algorithm's objective is to precisely detect action instances $\mathbf{P}_i = \{(t_j^{\mathrm{s}}, t_j^{\mathrm{e}}, c_j, s_j)\}_{j=1}^{N_i}$ within each untrimmed video. Specifically, the algorithm discovers N_i action instances in the i^{th} video, with the start time and end time of the j^{th} instance denoted as t_j^{s} and t_j^{e}, respectively. Additionally, c_j represents the predicted category, and s_j is the confidence score associated with this prediction.

It is essential to note that the distinction between the supervised paradigm and the weakly supervised paradigm lies in the nature of the supervision label $\mathbf{Y}$, leading to specific challenges for each setting. The supervised setting benefits from precise supervision, allowing for accurate action localization. Conversely, the weakly supervised setting poses challenges due to its reliance on video-level labels, leading to the need for innovative approaches to precisely identify action instances in untrimmed videos.

In the supervised setting, each training video is accompanied by a detailed supervision label, denoted as $\mathbf{Y}_i = \{(t_j^{\mathrm{s}}, t_j^{\mathrm{e}}, c_j)\}_{j=1}^{U_i}$. This label provides explicit information, including the start time, end time, and category label for each action instance within the video. Here, U_i denotes the total number of instances present in the video. The loss function, as represented in Eq.(1), takes into account both the classification and regression losses. Supervised algorithms leverage these annotations to learn action patterns comprehensively, albeit facing challenges stemming from significant variations in action duration [1], [14], [36], [42], [50], as well as discrepancies in action patterns [15], [16], [51], [52], and other factors.

Conversely, the weakly supervised setting involves training videos with video-level classification labels, denoted as $\mathbf{Y}_i = [c_i^1, c_i^2, ..., c_i^C]$, where $c_i^l = 1$ indicates the presence of action instances from the l^{th} category in the video. As video-level classification labels only reveal the existence of action instances without specifying instance number or temporal boundaries, the loss function in Eq.(1) primarily emphasizes video-level classification loss. Weakly supervised algorithms grapple with ambiguity issues arising from action-context confusion [53], [54], [55], [56], part domination [57], [58], [59], and false alarms in backgrounds [60].

2.2 History and scope

To the best of our knowledge, the origins of temporal action localization can be traced back to the concept of action search [61], [62], wherein algorithms aim to identify similar action instances within untrimmed videos, guided by an exemplary action instance. Zelnik et al. [61] propose a simple statistical distance measure for dynamic events and represented long-term temporal objects with spatiotemporal features at multiple temporal scales. Similarly, Gorelick et al. [62] treat human action as the silhouette of a moving torso and developed a 2D shape analyzing method to handle 3D spacetime actions. Building upon these initial works, Laptev et al. [63] and Duchenne et al. [64] introduced the concept of learning action patterns from movie scripts and localizing action boundaries, which marked the inception of weakly supervised temporal action localization. Laptev et al. [63] design a unified framework that simultaneously incorporates local space-time features, space-time pyramids, and multichannel non-linear SVMs for video classification, leveraging movie scripts to automatically discover human actions. In summary, traditional approaches for temporal action localization encompass diverse techniques, such as space-time features [62], spatiotemporal graphs [65], hidden Markov models [66], and more.

With the advent of deep learning research [67], [68], temporal action localization has undergone a revolutionary transformation. Leveraging the capabilities of video recognition backbones [69], [70], [71], which provide representative features, and adopting the end-to-end learning paradigm [36], which simplifies complex designs, the field has seen significant advancements. In the realm of supervised approaches, the anchor mechanism has seen notable developments, resulting in one-stage methods [33], [39], [72], [73], two-stage methods [14], [36], [52], [74], and anchor-free methods [44], [75], [76], [77]. On the other hand, in the context of weakly supervised methods, the community has introduced the pre-classification pipeline [2], [78], [79], [80] and the post-classification pipeline [20], [54], [81], [82].

It is essential to highlight that this survey centers specifically on temporal action localization research while considering other relevant tasks as distinct and separate topics. For instance, tasks such as action segmentation [83], spatiotemporal action detection [84], and action recognition [69], [70], [71] are not the primary focus of this survey and are treated as separate subjects.

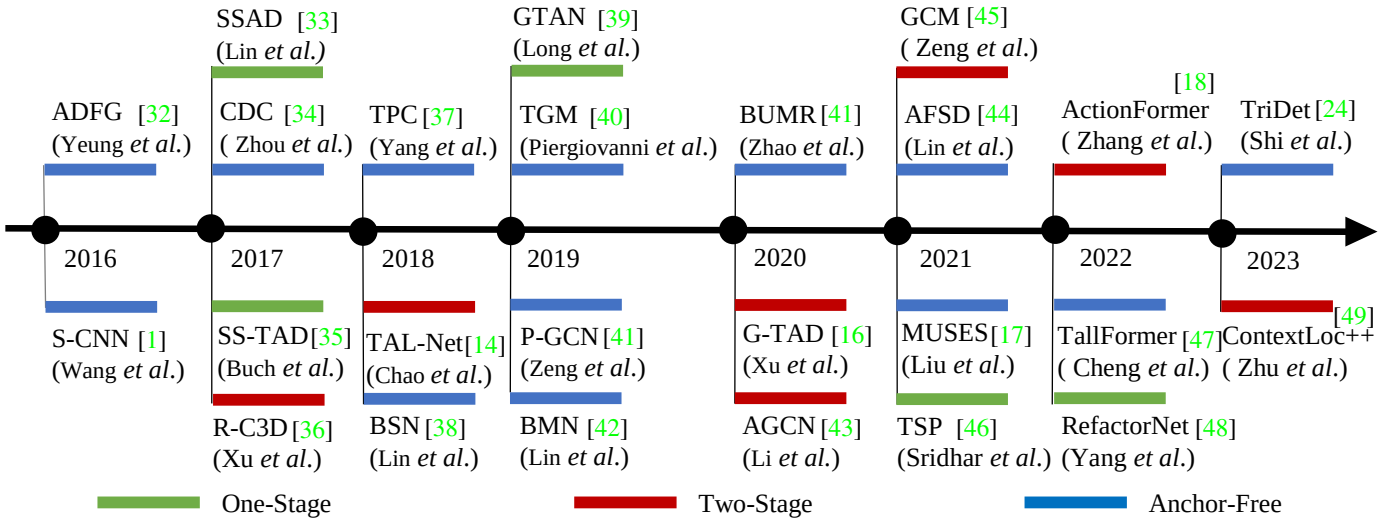

Fig. 3: Chronology for representative supervised temporal action localization research.

3 Fully Supervised Temporal Action Localization

In the deep learning era, temporal action localization methods can be classified into three main pipelines: the one-stage pipeline, the two-stage pipeline, and the anchor-free pipeline[1]. This categorization is based on whether the algorithm relies on default anchors and how these anchors are utilized. Furthermore, apart from the anchor mechanism taxonomy, we propose to review temporal action localization methods from the perspective of the classification mechanism. We categorize existing works into two distinct pipelines: the frame classification pipeline and the proposal classification pipeline. Additionally, we delve into two influential aspects of existing methods: the receptive field and feature representation. We summarize the representative characteristics of these aspects to gain valuable insights into their contributions to the field. Lastly, we identify the current bottleneck in supervised action localization algorithms and discuss several promising research directions aimed at overcoming these limitations.

As depicted in Fig. 3, several significant research works have emerged, each contributing to the advancement of action localization accuracy and serving as pivotal milestones in the field's chronology. In the realm of supervised temporal action localization, the earliest work can be traced back to detecting actions through the classification of sliding-window proposals [1]. Subsequently, Gao et al. [13] introduce the anchor mechanism to enhance proposal flexibility and accomplished action detection through regression based on default anchors. The pipeline's progress continued with the introduction of TAL-Net [14], which further developed the anchor mechanism into a two-stage approach. Subsequently, Zeng et al. [15] emphasize that action instances should not be detected individually and introduced graph convolution to model complex relationships among proposals, a concept further promoted by G-TAD [16]. Afterwards, Liu et al. [17] identify that sufficient comparisons among clustered proposals are key factors in modeling relationships among proposals. They transformed the task of localizing action scope in the temporal space into predicting a point in the 2D start-end space. More recently, ActionFormer [18] pushes the localization performance to a brand-new status, which is further developed by TriDet [24].

To ensure clarity in our review, we present an elaborate categorization of existing research into two main groups and five sub-groups, as detailed in Table 2. Furthermore, we provide corresponding schematic diagrams for each sub-group, visually depicted in Fig. 4. In the subsequent sections, we conduct a comprehensive review of the classification mechanism and the anchor mechanism. We discuss the procedures, representative works, strengths, and weaknesses associated with each mechanism.

3.1 Anchor mechanism

One-stage. The anchor mechanism has demonstrated remarkable efficiency in temporal action localization tasks, with existing works belonging to the one-stage pipeline, two-stage pipeline, or anchor-free pipeline. In particular, the one-stage pipeline [33], [37], [73], [85] stands as a simple yet effective solution, simultaneously predicting temporal boundaries and action categories for each instance, as illustrated in Fig. 4 (a). Notably, two widely used strategies contribute to enhancing localization accuracy within this pipeline: (1) estimating the overlap score (i.e., temporal intersection over union) enables precise ranking of candidate instances; (2) multiple pyramid layers facilitate the capture of action instances with varying durations. Additionally, the one-stage pipeline has witnessed significant advances, such as transforming the vanilla temporal convolutional kernel into a Gaussian kernel capable of modeling action structures [39], performing boundary regression and action classification via separate branches [72], and fusing appearance and motion features in a middle-level manner [86]. To summarize, the one-stage pipeline effectively discovers action instances. However, its performance is constrained by default anchors, necessitating careful tuning of hyperparameters to achieve optimal accuracy.

Two-stage. The two-stage pipeline operates by first generating action proposals using default temporal anchors and then performing elaborate boundary regression and category label prediction, as depicted in Fig. 4 (b). A representative work for the two-stage

1. We use "anchor-free" to indicate methods that do not rely on the anchor mechanism, which has a broader scope than the term's usage in the object detection domain.

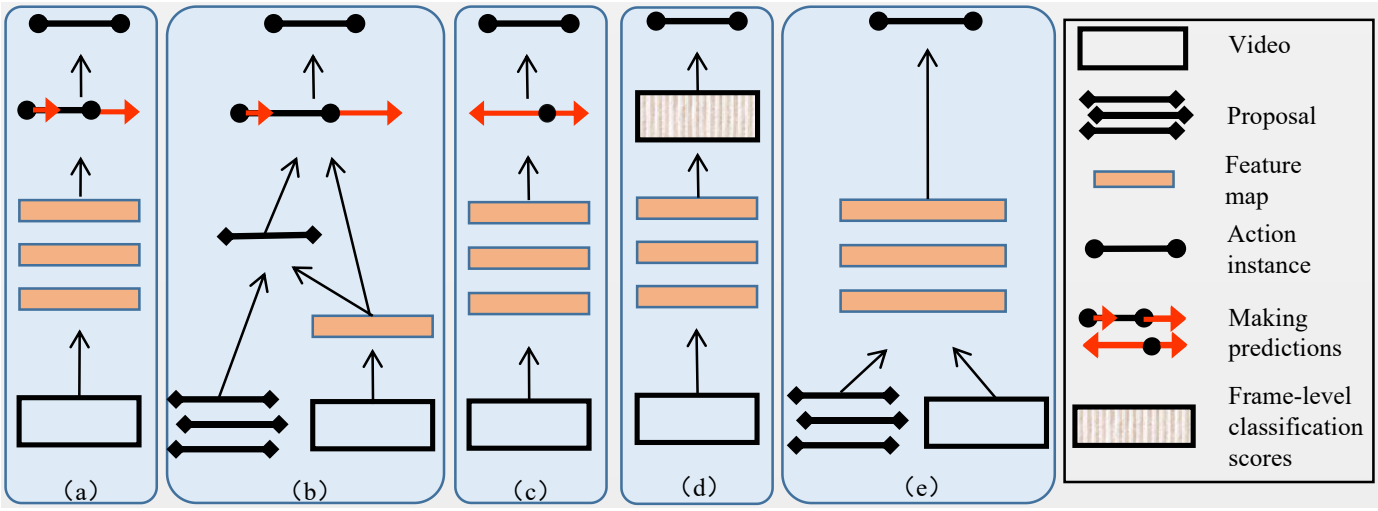

Fig. 4: Schematic diagram for five temporal action localization paradigms: (a) One-stage pipeline, (b) Two-stage pipeline, (c) Anchor-free pipeline, (d) Detection by frame classification, (e) Detection by proposal classification.

TABLE 2: Categorization of existing temporal action localization works. According to the operation mechanism, we group existing works into two categories, where each category can be further divided into sub-categories.

Paradigm	Category	Publications	Strengths	Weaknesses
Anchor Mechanism	One-Stage	[85], [33], [35], [37], [39], [72] [51], [77], [86], [87], [73], [88], [47]	• Efficient framework	• Parameter sensitive
	Two-Stage	[13], [36], [14], [74], [89], [43], [52], [90], [91]	• Accurate boundary • Elaborate fusion	• Complicated designs • Parameter sensitive
	Anchor-Free	[75], [76], [77], [44], [92], [93], [18], [94], [95], [18], [96]	• Simple design • Robustness	• Center uncertainty
Classification Mechanism	Classifying frames	[32], [97], [34], [98], [38], [37], [99], [40], [51], [41], [100], [50], [43], [101], [102], [88], [46], [103], [92], [18], [104], [103]	• Accurate boundary • Tackling long instances	• Lacking temporal modeling • Separated procedures
	Classifying proposals	[1], [105], [13], [106], [98], [42], [15], [107], [108], [109], [110], [17], [52], [45], [101], [90], [49]	• High recall • Low background error	• Heavy computation • Not flexible enough

pipeline is R-C3D [36], which has been further developed by TAL-Net [14]. Given its multiple procedures, the two-stage pipeline offers opportunities for specific designs tailored to temporal action localization tasks. For instance, the receptive field alignment module [14] addresses drastic variations in action instances, while the attention-based graph convolutional module [43] explores intra- and inter-proposal relationships. On one hand, the two-stage pipeline has the potential to achieve accurate action boundaries. On the other hand, compared to the one-stage pipeline, it is more sensitive to hyperparameters and necessitates more intricate designs.

Anchor-free. In contrast to the one-stage and two-stage pipelines that rely on default anchors, the anchor-free pipeline directly performs action boundary regression without referencing default anchors, as illustrated in Fig. 4 (c). While not extensively emphasized, credit for pioneering work goes to [75], which predicted temporal distances from each frame to potential action boundaries. Building on this foundation, Tang et al. [76] draw inspiration from FCOS [121] and introduce an anchor-free temporal action detection framework, known as AFO-TAD. Subsequently, Yang et al. [77] conduct an in-depth study on the anchor-free pipeline, revealing its superiority in handling extremely long or extremely short action instances. More recently, Lin et al. [44] further advance the anchor-free pipeline by learning salient boundary features and proposing a prediction-and-refinement framework. In summary, the anchor-free pipeline exhibits robustness and simplicity in design, but it may encounter challenges related to the center uncertainty problem.

3.2 Classification mechanism

Frame classification. Temporal action localization necessitates effective spatio-temporal modeling. In the frame classification paradigm, spatial information is initially modeled by classifying each frame, followed by perceiving temporal evolution through pre-defined rules, as depicted in Fig. 4 (d). For instance, some works [34], [40], [97] set a threshold on each frame's classification score and consider consecutive frames whose scores exceed the threshold as an action instance. Additionally, other approaches [38], [41], [50], [99] estimate the likelihood of start and end moments for each frame and associate potential start and end moments to form action instances. In summary, the frame classification pipeline can achieve accurate action boundaries and effectively discover long-duration action instances. However, addressing each frame individually poses a limitation in temporal modeling, often leading to false alarms in backgrounds. Moreover,

TABLE 3: Summary of existing temporal action localization methods from two important methodology views: receptive field and feature representation.

Methodology	Category	Publications	Strengths	Weaknesses
Receptive Field	Intra-video local relationship	[34], [33], [13], [36], [33], [37], [14], [76], [77], [38], [42], [40], [51], [39], [75], [41], [52], [44], [103], [92], [93], [92], [104], [103]	• Detailed features	• Lacking globality • Insufficient intra-video diversity
	Intra-video global relationship	[1], [97], [35], [106], [99], [15], [110], [43], [45], [111], [90], [46], [104], [112], [88], [113], [18], [102], [94], [95], [47], [18], [49], [91], [96]	• Temporal modeling • Intra-video diversity	• Insufficient detail representation
	Inter-video relationship	[114], [115]	• Representative category features	• Complicated training
Feature Representation	End-to-End	[1], [34], [105], [36], [114], [116], [108], [44], [73], [117], [113], [93], [47], [96]	• End-to-end learnable • Representative features	• Huge GPU footprint • Heavy computation
	Pre-extracting	[118], [13], [33], [14], [38], [15], [39], [42], [107], [86], [87], [16], [119], [109], [45], [88], [90], [46], [112], [103], [92], [18], [102], [92], [104], [94], [103], [95], [18], [49], [91]	• Efficient to train and inference	• Unflexible feature representation
	Finetuning and Pre-extracting	[120], [40], [52], [17]	• Trade-off between computation and representation	• Inconvenience of deployment

the reliance on multiple separated procedures, governed by hand-crafted rules, may constrain the pipeline's performance.

Proposal classification. In addition to the frame classification pipeline, the proposal classification pipeline has garnered increasing research interest in recent years, as illustrated in Fig. 4 (e). This pipeline revolves around two fundamental problems: (1) obtaining the proposal, and (2) performing classification. Regarding the first problem, pioneering works [1], [122] have employed the sliding window strategy to construct action proposals. Subsequently, researchers have introduced the temporal actions grouping strategy [105], [107], [123] and the dense enumeration strategy [17], [42]. As for the second problem, early explorations [1], [122] independently performed classification and regression for each action proposal. However, the community has recognized the impact of relationships among multiple proposals and subsequently developed the graph convolutional theory to simultaneously consider multiple proposals [15], [45], [107]. By leveraging proposals, the proposal classification pipeline can achieve a high recall rate and effectively reduce background errors. However, a large number of proposals may result in heavy computational overhead, and the use of fixed proposals can limit the algorithm's flexibility.

3.3 Categorization from the methodology view

This section provides a comprehensive review of existing temporal action localization methods from a methodological perspective. We focus on two crucial factors, namely, the receptive field and feature representation, and analyze their influence on the localization performance. The detailed categorization of the reviewed methods is presented in Table 3.

Receptive field. The receptive field plays a crucial role in accurately determining action boundaries, as effective temporal modeling is essential for temporal action localization. In the early explorations [38], [98], [123], video frames were independently processed to learn feature representation and perform classification, resulting in limited temporal receptive fields. To address this limitation, the pyramid network architecture [16], [33], [39], [86] was introduced, which employs multiple layers to gradually increase the receptive field layer-by-layer. Although these approaches primarily capture intra-video local relationships and can learn detailed feature representations from each frame, they are constrained by their local receptive fields. As a result, these pipelines may struggle to model global relationships and are limited in their ability to jointly learn from multiple instances within a video.

To enhance the temporal receptive field, the community has proposed three types of solutions: recurrent memory, graph models, and transformers. In the case of recurrent memory, early attempts by Escorcia et al. [122] and Ma et al. [97] employ Long Short-Term Memory (LSTM) [126] to capture temporal dependencies. This approach was further developed by others [35], [99], [106] with the introduction of Gated Recurrent Unit (GRU) [127]. On the other hand, for graph models, P-GCN [45] used graph convolutions to capture relationships among action proposals, while G-TAD [16] proposes the sub-graph alignment layer to discover action instances by classifying sub-graphs. More recently, transformer architectures, known for their effectiveness in modeling long-term dependencies, have been adapted for action localization tasks. Early attempts include the relaxed transformer decoder network [111] and augmented transformer [101]. These methods offer effective temporal modeling and the ability to associate multiple instances within an untrimmed video, thus addressing the intra-video diversity. However, their focus on globality modeling may lead to insufficient detail representation and hinder boundary accuracy.

In addition to exploring intra-video relationships, there have been some works [114], [115] that investigate inter-video relationships. Wang et al. [114] propose a siamese network to simultaneously process two proposals and verify their similarity. On the other hand, Zhang et al. [115] associate co-occurring action instances within two videos using a cross-video similarity matrix. The modeling of inter-video relationships contributes to learning representative features for each action category. However, it is important to note that such an inter-video pipeline cannot independently localize actions and often collaborates with traditional intra-video methods.

Feature representation. The choice of feature representation directly impacts the performance of action localization. Early works adopted off-the-shelf video recognition backbones for end-

TABLE 4: Summary of remarkable characteristics of supervised temporal action localization methods.

Methods	Publications	Feature Extractor	End-to-End	Flow	Temporal Procedure	Proposal	Performance (%)	
							THUMOS14	ANet v1.3
Heilborn et al. [118]	CVPR 2016	STIPs		✓	Multi-scale sliding window	No	13.5	-
DAPs [122]	ECCV 2016	C3D	✓		Single-scale sliding window	DAPs [122]	13.9	-
Yeung et al. [32]	CVPR 2016	VGG-16	✓		Single-scale sliding window	No	17.1	-
S-CNN [1]	CVPR 2016	C3D	✓		Multi-scale sliding window	No	19.0	-
SST [106]	CVPR 2017	C3D	✓		Single-scale sliding window	SST [106]	23.0	-
CDC [34]	CVPR 2017	C3D	✓		Single-scale sliding window	No	23.3	23.80
SSAD [33]	ACM 2017	C3D + TS		✓	Single-scale sliding window	No	24.6	-
R-C3D [36]	ICCV 2017	C3D	✓		Single-scale sliding window	No	28.9	12.70
SSN [105]	ICCV 2017	TS	✓		-	TAG [123]	29.1	28.28
Buch et al. [35]	BMVC 2017	C3D	✓		Single-scale sliding window	No	29.2	-
CBR [13]	BMVC 2017	TS		✓	-	TURN [85]	31.0	-
CTAP [98]	ECCV 2018	TS		✓	Multi-scale sliding window	No	29.9	-
Wang et al. [124]	ICPR 2018	C3D	✓		Multi-scale sliding window	No	32.2	-
BSN [38]	ECCV 2018	TS		✓	Single-scale sliding window	No	36.9	33.72
TAL-Net [14]	CVPR 2018	I3D		✓	Vanilla length	No	42.8	20.22
GTAN [39]	CVPR 2019	P3D			Single-scale sliding window	No	38.8	34.31
BMN [42]	ICCV 2019	TS		✓	Single-scale sliding window	No	38.8	33.85
PGCN [15]	ICCV 2019	I3D		✓	-	BSN [38]	49.1	31.11
Zhou et al. [108]	TMM 2020	TS	✓	✓	-	TAG [123]	42.6	-
Zhao et al. [41]	ECCV 2020	I3D		✓	Fixed Length	No	45.4	30.12
A2Net [77]	TIP 2020	I3D		✓	Single-scale sliding window	No	45.5	27.75
AFNet [89]	TMM 2020	C3D	✓		Multi-scale sliding window	No	49.5	18.60
PBRNet [50]	AAAI 2020	I3D		✓	Single-scale sliding window	No	51.3	35.01
G-TAD [16]	CVPR 2020	I3D		✓	Single-scale sliding window	No	51.6	34.09
C-TCN [87]	ACM 2020	I3D		✓	Single-scale sliding window	No	52.1	31.10
BSN++ [109]	AAAI 2021	TS		✓	Single-scale sliding window	No	41.3	34.88
TCANet [52]	CVPR 2021	Slowfast [FT]		✓	-	BMN [42]	44.6	35.52
PCG-TAL [119]	TIP 2021	I3D		✓	Single-scale sliding window	No	51.2	28.85
GCM [45]	TPAMI 2021	I3D		✓	-	BSN [38]	51.9	34.24
ContextLoc [90]	ICCV 2021	I3D		✓	-	BSN [38]	54.3	34.23
AFSD [44]	CVPR 2021	I3D [FT]	✓		Single-scale sliding window	No	55.5	34.40
MUSES [17]	CVPR 2021	I3D [FT]		✓	-	BSN [38]	56.9	33.99
RCL [92]	CVPR 2022	I3D		✓	Single-scale sliding window	No	52.9	37.65
DCAN [104]	AAAI 2022	I3D		✓	Single-scale sliding window	No	54.1	35.39
TAGS [125]	ECCV 2022	TS		✓	Single-scale sliding window	No	57.0	36.50
ReAct [94]	ECCV 2022	I3D		✓	Single-scale sliding window	No	57.1	32.60
RefactorNet [103]	CVPR 2022	I3D		✓	Single-scale sliding window	No	58.6	38.60
TadTR [95]	TIP 2022	I3D		✓	Single-scale sliding window	No	60.1	36.75
TallFormer [47]	ECCV 2022	Swin [FT]	✓		Single-scale sliding window	No	63.2	35.60
ActionFormer [18]	ECCV 2022	I3D		✓	Single-scale sliding window	No	71.0	36.60
ContextLoc++ [49]	PAMI 2023	I3D		✓	Single-scale sliding window	BSN [38]	58.7	38.13
SoLa [91]	CVPR 2023	TS		✓	Single-scale sliding window	No	59.1	34.99
Re^{2}TAL [96]	CVPR 2023	Slowfast [FT]	✓		Single-scale sliding window	No	64.9	37.01
TriDet [24]	CVPR 2023	I3D		✓	Single-scale sliding window	No	72.9	36.80

For the column *Feature Extractor*, [FT] indicates finetuning the video recognition model on temporal action localization dataset.

to-end learning. Some widely used backbones include the C3D network [69], utilized by [1], [34], [36], and the TSN network [128], employed by [105], due to their ability to extract representative spatiotemporal features [129]. Additionally, certain works [73], [97] first process each frame individually using an image recognition backbone (e.g., VGG [68], ResNet [130]), and then capture temporal evolution through recurrent strategies [97] or non-local strategies [73]. This end-to-end learning approach allows adaptive learning of representative features for the temporal action localization task, but it necessitates significant GPU memory and computational resources.

To alleviate the computation burden, many researchers adopt a two-step approach: first, they pre-extract features from untrimmed input videos, and then they train the action localization model. This pipeline is widely used due to its efficiency in training. In this context, the I3D [70] network, pre-trained on the Kinetics dataset [70], and the two-stream network [128], pre-trained on the ActivityNet dataset [131], emerge as two dominant backbones. To improve the feature representation, some studies incorporate additional fully-connected or convolutional layers to project the vanilla features into a feature space more suitable for the action localization task. However, despite these efforts, the feature representation still lacks flexibility, which limits the localization performance as it heavily depends on the recognition model used for feature extraction.

To allevaite this issue, recent works [17], [44], [52] propose a two-step approach, involving finetuning the backbone network and then extracting features. Additionally, efforts have been devoted to reducing memory consumption and enabling end-to-end training [96], [141]. Notably, TSP [141] enhances temporal sensitivity by simultaneously considering action-level classification and temporal-region classification. The TSP features have become rep-

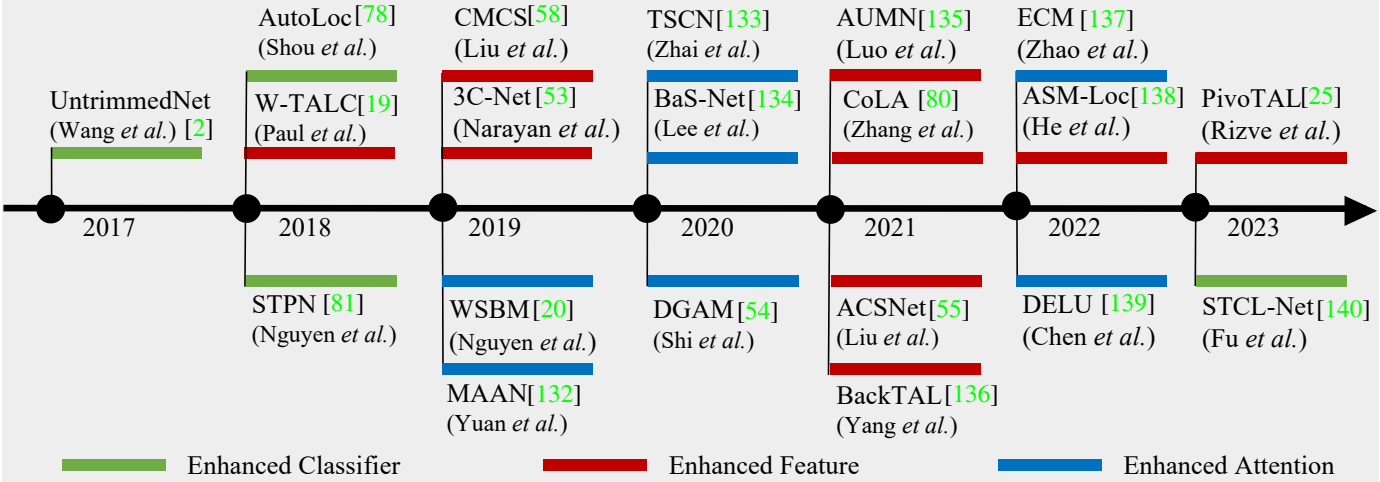

Fig. 5: Chronology for weakly supervised temporal action localization research.

resentative and widely adopted in subsequent works, particularly in action localization challenges. More recently, Re^{2}TAL [96] introduces a reversible module that clears intermediate activations during training, thereby mitigating the memory footprint bottleneck. The combination of backbone finetuning and efficient end-to-end training strikes a favorable balance between computational burden and feature representation, leading to performance gains in the aforementioned works.

Characteristic summary and analyses. Table 4 presents a comprehensive summary of existing temporal action localization methods, revealing three prominent trends. (1) The evolution of video recognition backbones has significantly benefited action localization research. In the deep learning era, representative backbones such as the C3D network [69], two-stream network [2], [128], [142], and I3D network [70] have played pivotal roles. Recent works have explored methods to access more representative features, either through finetuning the I3D network or adopting the newly proposed Slowfast network [71]. Besides capturing appearance features from video frames, most approaches rely on optical flow to describe motion characteristics. Additionally, pre-extraction of features is a common practice, streamlining the training and inference processes. (2) The sliding window strategy has proven to be an effective approach for handling untrimmed videos. Earlier works [1], [98], [118] employed multiple sliding windows with different temporal scales to segment untrimmed videos. However, subsequent research [33], [34], [36] demonstrated that a single-scale sliding window is often sufficient to localize action instances accurately. Furthermore, Chao et al. [14] retained the original length of each video, while Zhao et al. [41] scaled untrimmed videos to a fixed length for consistency. (3) While there have been significant performance gains on the THUMOS14 benchmark, the improvement on ActivityNet v1.3 has been comparatively slow. This observation may be attributed to two key factors. First, ActivityNet v1.3 employs a stricter evaluation metric, the average mAP under threshold [0.5:0.05:0.95], posing a higher bar for achieving top performance. In contrast, THUMOS14 uses the relatively looser metric, mAP under threshold 0.5. Second, ActivityNet v1.3 presents greater difficulty due to a higher number of categories and larger variation in action durations compared to THUMOS14.

3.4 Further discussions and promising directions

Temporal action localization is a foundational task in intelligent video understanding, and it is essential to develop theoretical research while considering practical applications. In recent work, Damen et al. [143] introduced a method for localizing action instances in egocentric videos, which finds applications in diverse scenarios such as robotics and industrial manufacturing. However, algorithms designed for egocentric videos must address challenges related to the presence of severe noise in optical flow inputs. The camera's movements inevitably introduce significant background motions, necessitating robust algorithms to handle such complexities. Another practical area of interest is online action detection [144], [145]. This topic is particularly relevant in applications like smart surveillance, where algorithms are expected to trigger alarms promptly when an anomaly event starts, rather than merely localizing it in an offline manner. To achieve swift detection, algorithms need to extract sensitive boundary features from noisy input data and strike a balance between computational efficiency and feature representation capability. In conclusion, the temporal action localization community should not only focus on theoretical advancements but also keep in mind the real-world applications where the developed methods will be deployed. Addressing the challenges posed by specific scenarios, such as egocentric videos and online action detection, will lead to more practical and effective solutions.

Both egocentric action localization and online action localization present significant challenges, but recent progress in related areas can serve as valuable inspiration for further explorations. For instance, Wang et al. [73] have demonstrated promising results by employing only RGB frames to construct an end-to-end temporal action localization baseline. This accomplishment highlights two key points: (1) With appropriate designs, a sequence of RGB frames can effectively capture motion characteristics, eliminating the need for optical flow inputs. (2) The end-to-end learning pipeline proves to be capable of dynamically extracting representative features without imposing excessive computational burdens.

In conclusion, we encourage future research to continue advancing the field of temporal action localization with a practical mindset. This entails addressing specific challenges such as localizing actions in egocentric videos and enabling online action detection. By considering real-world applications and developing

effective solutions, the temporal action localization community can make valuable contributions to the field of intelligent video understanding.

4 Weakly Supervised Temporal Action Localization

In the deep learning era, weakly supervised temporal action localization research can be categorized into two groups: pre-classification research and post-classification research. Pre-classification involves performing classification on snippet-level features, while post-classification is based on aggregated features. Researchers have proposed various strategies to enhance feature extraction, attention, and classifiers. In this section, we analyze the characteristics of existing works and provide suggestions for future research directions.

Fig. 5 presents a concise chronology of weakly supervised temporal action localization methods. UntrimmedNet [2] is an early pioneering work that employs effective procedures such as predicting video-level classification scores using top-k mean strategy and detecting actions via thresholding. Subsequently, W-TALC [19] extends the pipeline to an inter-video manner, considering two videos from the same category to learn representative features. Nguyen et al. [20] address the critical influence of background segments on weakly supervised learning, which was overlooked by previous research, and propose the background modeling pipeline. EM-MIL [21] identifies that early fusion and late fusion, utilized in previous works, do not fully explore the complementarity between appearance and motion features, and they propose cross-branch supervision to mine this complementarity. More recently, Liu et al. [151] view background segments as blessings rather than curses, integrating the causal analysis pipeline to enhance localization accuracy.

4.1 Classification mechanism

Under the weakly supervised setting, the learning process is constrained by the availability of only video-level classification labels, while precise localization of action instances necessitates frame-level predictions. The association between frame-level predictions and video-level labels is not unique, presenting challenges to weakly supervised learning algorithms.

Pre-classification. As depicted in Fig. 6 (a), the pre-classification pipeline begins by utilizing several temporal convolutional layers to classify each frame. It then predicts the video-level classification label by aggregating a series of frame-level scores, achieved through methods like weighted sum [2], [148], [149] or the top-k mean strategy [53], [59], [134]. Leveraging temporal convolutional operations allows this pipeline to generate discriminative local features and effectively avoid noise interference in background segments. However, the limited receptive field of the temporal convolutional layers poses a challenge, as it restricts the pipeline's ability to model long-term relationships. As a result, pre-classification methods tend to identify the most discriminative action parts but often struggle to generate complete action instances, resulting in what is referred to as the "part domination problem."

To alleviate the challenge of part domination, the research community has proposed various solutions, which can be classified into four groups: hiding discriminative parts, pseudo label training, learning multiple branches, and prior-based designs. Firstly, the

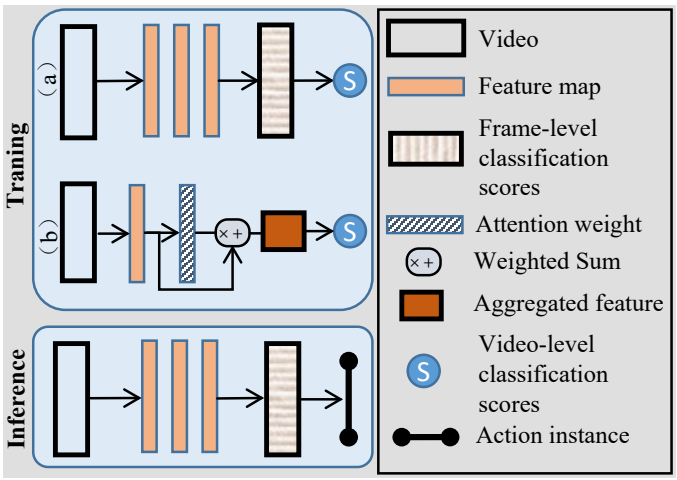

Fig. 6: Schematic diagram for (a) pre-classification mechanism and (b) post-classification mechanism, where both training and inference pipelines are shown.

"hide-and-seek" approach [57], along with its extension [146], intentionally conceals discriminative action parts during training. This encourages the algorithm to explore inconspicuous regions and generate complete action instances. Additionally, some weakly supervised learning algorithms [21], [78], [79], [147] adopt a pseudo label training strategy. They generate pseudo labels for video frames and then train the localization network under a supervised manner. This approach assumes that multiple pseudo instances may contain complementary cues, leading to the identification of complete action patterns. Moreover, recognizing that diverse classification scores might form complete action instances, Liu et al. [58] propose to learn multiple branches in parallel, requiring the classification responses to exhibit diversity. This approach has been further improved through adversarial training [59]. Furthermore, some methods leverage priors to enhance the localization performance. For instance, Lee et al. [150] propose that background features should have small magnitudes, while action features should have large magnitudes. Others decompose action instances into multiple subactions [23], [135], [186] to better exploit prior knowledge.

Despite the progress made by the four pipelines in alleviating the part domination problem, the pre-classification approach still encounters certain challenges. In the hiding discriminative parts pipeline, it remains uncertain when to cease the hiding process, as it is difficult to ascertain if an action instance has been completely discovered. Additionally, the performance of the pseudo label training pipeline is often limited by the quality of generated pseudo action instances, which can affect the overall localization accuracy. Moreover, the learning multiple branches pipeline necessitates careful consideration of the trade-off between the classification accuracy of each branch and the diversity of responses among multiple branches. Lastly, the prior-based design pipeline is relatively inflexible and may require intricate designs to generalize effectively to new scenarios.

Post-classification. As depicted in Fig. 6 (b), the post-classification pipeline adopts a two-step approach. Initially, it evaluates the relevance of each frame to the video-level classification task and assigns an attention weight accordingly. Subsequently, the vanilla video features are aggregated into a feature vector, which is utilized to train the classifier and predict the video-level

TABLE 5: Categorization of existing weakly supervised temporal action localization works. According to the operation mechanism, we group existing methods into two categories, where each category can be further divided into sub-categories.

Paradigm	Category	Publications	Strengths	Weaknesses
Classification Mechanism	Pre-Classification	[2], [146], [78], [79], [58], [53], [147], [148], [59], [133], [21], [149], [150], [135], [80], [151], [152], [134], [22], [153], [154], [155], [156], [157], [158], [159], [160], [161], [162], [163], [138], [164], [165], [166], [167], [168], [139], [169], [25], [170], [140], [171], [172], [173], [174]	• Discriminative local features • Less background interference	• Lacking long-term modeling • Dominated by action part
	Post-Classification	[57], [81], [175], [20], [132], [54], [55], [56], [23], [82], [176], [177], [178], [179]	• Global awareness • Associating multiple instances	• Insufficient detail modeling • Action-context confusing
Enhancement Mechanism	Enhanced Classifier	[57], [146], [81], [78] , [79], [20], [59], [147], [134], [159], [160], [174], [140]	• Alleviating part domination and action-context confusion	• Requiring specific designs
	Enhanced Feature	[180], [58], [53], [55], [56], [150], [135], [80], [152], [23], [82], [156], [181], [155], [177], [157], [158], [178], [161], [163], [182], [138], [164], [179], [165], [167], [183], [173], [172], [171], [170]	• Intra-class similarity • Inter-class discrepancy	• Relying on priors • Hyper-parameter sensitive
	Enhanced Attention	[175], [132], [148], [54], [133], [149], [22], [153], [154], [176], [166], [168], [184], [185], [169], [25]	• A simplification of localizing actions	• Relying on proper normalization

classification score. During inference, the classifier is applied to classify each frame and identify action instances, as illustrated in the lower part of Fig. 6. Since the classifier learns from aggregated features, the algorithm can effectively capture global cues in untrimmed videos and associate multiple action instances within the video. However, the use of aggregated features may result in insufficient modeling of local details, leading to challenges in distinguishing between boundary action frames (e.g., the end of a diving action) and neighboring background frames (e.g., water splashing after diving). This phenomenon is known as the action-context confusion challenge.

To address the action-context confusion challenge, the research community has proposed three types of solutions: attention modeling, background modeling, and context modeling. The attention modeling pipeline [54], [132] focuses on generating high-quality attention maps to effectively separate action frames from neighboring background frames. For instance, Yuan et al. [132] introduce a marginalized average aggregation module to suppress the response of discriminative regions, while Shi et al. [54] incorporate representability into the estimation of attention weights. On the other hand, the background modeling pipeline [20], [59], [148] emphasizes the importance of understanding background information. Nguyen et al. [20] propose to estimate background attention by inverting action attention, followed by aggregating background features and predicting background-aware classification scores. This strategy is further extended and refined by subsequent works [59], [148]. Additionally, the context modeling pipeline [55], [56] considers that foreground frames consist of both action frames and contextual frames, while background frames are the opposite of foreground frames. To distinguish between action and context, ACSNet [55] combines latent components and performs action-context classification. Similarly, Liu et al. [56] learn explicit subspaces for action and context separately. Although the above pipelines have made progress in alleviating the action-context confusion, it remains a challenging issue, especially in scenarios where the algorithm has access to only video-level classification labels during training. Further research and improvements are needed to fully address this challenge.

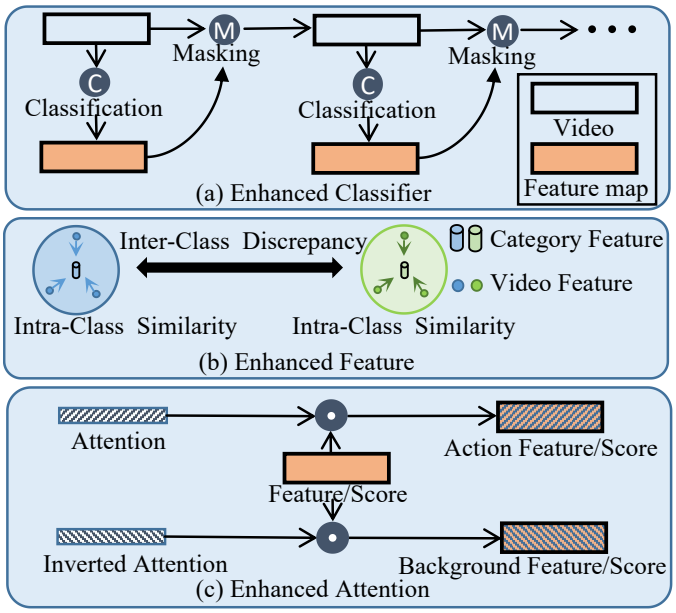

Fig. 7: Schematic diagram for (a) enhanced classifier mechanism, (b) enhanced feature mechanism and (c) enhanced attention mechanism.

4.2 Enhancement mechanism

In addition to the classification mechanism, weakly supervised temporal action localization methods can be categorized based on the enhancement mechanism. As shown in Table 5, we analyze the enhancement mechanism from three aspects: enhanced classifier, enhanced feature, and enhanced attention.

Enhanced Classifier. Because weakly supervised learning algorithms can only access video-level classification labels, an enhanced classifier is beneficial for localizing action instances, as shown in Fig. 7 (a). Singh et al. [57] and Zhong et al. [146]

TABLE 6: Summary of remarkable characteristics of weakly supervised temporal action localization methods.

Methods	Publications	Feature Extractor	Supervision	Performance (%)		
				THUMOS14	ANet v1.2	ANet v1.3
Hide-and-Seek [57]	ICCV 2017	C3D	Cls. label + UCF101	6.8	-	-
UntrimmedNets [2]	CVPR 2017	End-to-End	Cls. label	13.7	-	-
STPN [81]	CVPR 2018	I3D	Cls. label	16.9	-	-
AutoLoc [78]	ECCV 2018	UNT	Cls. label	21.2	16.0	-
W-TALC [19]	ECCV 2018	I3D	Cls. label	22.8	18.0	-
MAAN [132]	ICLR 2019	I3D	Cls. label	20.3	-	-
STAR [175]	AAAI 2019	I3D	Cls. label	23.0	-	-
CMCS [58]	CVPR 2019	I3D	Cls. label	23.1	22.4	21.2
CleanNet [147]	ICCV 2019	UNT	Cls. label	23.9	21.6	-
TSM [187]	ICCV 2019	I3D	Cls. label	24.5	17.1	-
ASSG [79]	ACM MM 2019	I3D	Cls. label	25.4	-	-
3C-Net [53]	ICCV 2019	I3D	Cls. label + Action Count	26.6	21.7	-
WSBM [20]	ICCV 2019	I3D	Cls. label + Micro videos	27.5	-	-
PreTrimNet [188]	AAAI 2020	I3D	Cls. label	23.1	-	22.5
AdapNet [189]	TNNLS 2020	ResNet-101	Cls. label + UCF 101	23.7	-	22.0
BaS-Net [134]	AAAI 2020	I3D	Cls. label	27.0	24.3	22.2
TSCN [133]	ECCV 2020	I3D	Cls. label	28.7	23.6	21.7
DGAM [54]	CVPR 2020	I3D	Cls. label	28.8	24.4	-
A2CL-PT [59]	ECCV 2020	I3D	Cls. label + UCF 101	30.1	-	22.5
ACL [190]	CVPR 2020	I3D	Cls. label	30.1	24.6	-
EM-MIL [21]	ECCV 2020	I3D	Cls. label	30.5	20.3	-
Huang et al. [156]	TPAMI 2021	I3D	Cls. label	29.0	24.2	22.7
Liu et al. [56]	AAAI 2021	I3D	Cls. label	29.6	25.5	23.2
MSA-Net [82]	TIP 2021	I3D	Cls. label	29.7	25.7	23.5
ASL [152]	CVPR 2021	I3D	Cls. label	31.1	25.8	-
CoLA [80]	CVPR 2021	I3D	Cls. label	32.2	26.1	-
ACSNet [55]	AAAI 2021	I3D	Cls. label	32.4	26.0	23.9
HAM-Net [149]	AAAI 2021	I3D	Cls. label	32.6	25.1	-
AUMN [135]	CVPR 2021	I3D	Cls. label	33.3	25.5	23.5
UGCT [176]	CVPR 2021	I3D	Cls. label	33.3	25.8	23.8
Lee et al. [150]	AAAI 2021	I3D	Cls. label	33.7	25.9	23.7
TS-PCA [151]	CVPR 2021	I3D	Cls. label	34.3	-	23.7
D2-Net [155]	ICCV 2021	I3D	Cls. label	35.9	26.0	-
CO_2-Net [22]	ACM MM 2021	I3D	Cls. label	38.3	26.4	-
A-TSCN [185]	TPAMI 2022	I3D	Cls. Label	33.6	25.6	23.6
FTCL [179]	CVPR 2022	I3D	Cls. label	35.6	-	24.8
DCC [164]	CVPR 2022	I3D	Cls. label	35.7	-	24.3
MMSD [167]	TIP 2022	I3D	Cls. label	36.4	-	25.8
ASM-Loc [138]	CVPR 2022	I3D	Cls. label	36.6	-	25.1
Huang et al. [165]	CVPR 2022	I3D	Cls. label	38.2	-	25.0
ECM [137]	TPAMI 2022	I3D	Cls. label	29.1	25.5	23.5
DELU [139]	ECCV 2022	I3D	Cls. Label	40.5	26.9	-
AICL [174]	AAAI 2023	I3D	Cls. Label	36.9	29.9	27.6
Wang et al. [173]	CVPR 2023	I3D	Cls. label	37.1	-	26.3
Li et al. [172]	CVPR 2023	I3D	Cls. label	39.3	-	26.0
Ren et al. [171]	CVPR 2023	I3D	Cls. label	40.0	26.5	25.5
STCL-Net [140]	TPAMI 2023	I3D	Cls. label	41.8	26.6	24.7
Ju et al. [170]	CVPR 2023	I3D	Cls. label	42.0	29.6	-
Zhou et al. [169]	CVPR 2023	I3D	Cls. label	42.7	-	28.8
PivoTAL [25]	CVPR 2023	I3D	Cls. Label	42.8	-	28.1

For the *Supervision* column, "Cls. label" indicates video-level classification label.

remove the most discriminative action parts, and the classifier gets enhanced when highlighting other action parts. Lee et al. [134] and Nguyen et al. [20] employ the same classifier to dispose of both foreground features and background features, where the classifier gets enhanced via clearly distinguishing actions from backgrounds. AutoLoc [78] and CleanNet [147] first generate pseudo action instances, then enhance the classifier through training under a supervised manner. To sum up, the enhanced classifier mechanism has made progress in solving both the discriminative part dominant problem and the action-context confusion problem. However, researchers should make specific designs according to the enhancement target, which constrains the generalization of this pipeline.

Enhanced Feature. As most weakly supervised temporal action localization works rely on pre-extracted features, the feature enhancement pipeline aims to transform vanilla features into features suitable for action localization, as shown in Fig. 7 (b). Some works [53], [55], [56], [80], [156] enhance features by learning similar representations for instances from the same category and keeping them distinct for instances from different categories. In other words, they pursue both intra-class similarity and inter-class discrepancy. For example, Huang et al. [156] learn prototypes for

each category, while Liu et al. [55], [56] embed category features into distinct subspaces. Furthermore, some works transport features through several convolutional layers and directly impose constraints on features to achieve accurate action localization. For instance, they may focus on the magnitude of features [150] or the diversity of features [58]. Later, [23], [135], [186] observed the diversity within each category and proposed to learn subaction representations for each category, resulting in a more elaborate feature enhancement approach. In summary, the feature enhancement pipeline can generate high-quality features by preserving intra-class similarity and inter-class discrepancy. However, it relies on priors and is sensitive to hyper-parameters, which should be taken into consideration during the design process.

Enhanced Attention. The attention mechanism can guide the algorithm to focus on informative action parts while avoiding interference from background parts, as shown in Fig. 7 (c). In action localization research, some works introduce additional loss terms to learn reliable attention weights. For example, BANet [148] employs the action-guided attention loss, TSCN [133] utilizes the attention normalization loss, and DGAM [54] adopts a VAE module and calculates the reconstruction loss. Moreover, some works dynamically adjust the attention weights. For instance, Xu et al. [175] calculate class-variable attention weights, while Yuan et al. [132] dynamically select responses based on randomly sampled attention weights following a Bernoulli distribution. More recently, HAM-Net [149] utilizes three variants of the vanilla attention weight, namely semi-soft attention, soft attention, and hard attention, to enhance the attention weight. In general, determining attention weights can be seen as the simplification of localizing action instances, where the algorithm focuses solely on estimating action boundaries and does not predict the action category. However, the attention mechanism should be equipped with a proper normalization strategy; otherwise, degenerated attention weights could lead to a drop in performance.

4.3 Characteristics analyses

Table 6 provides a summary of remarkable characteristics observed in representative temporal action localization methods. Notably, most works maintain consistent experimental setups, facilitating fair comparisons among different algorithms. Regarding feature representation, the majority of methods utilize the I3D network [70], pre-trained on the Kinetics 400 dataset, to extract features from untrimmed videos. This approach captures both appearance and motion features, which are essential for accurately depicting untrimmed videos. However, it is important to note that networks such as I3D and [2], initially trained for video recognition tasks, may not offer flexible feature representations perfectly suited to the temporal action localization task due to the domain gap. In terms of supervision, most works solely rely on video-level classification labels. Some methods have explored incorporating additional supervision signals [20], [53] or pre-training the network with extra data from the UCF 101 dataset [191] [57], [59], [189]. However, it is noteworthy that learning from video-level classification labels remains the dominant trend and achieves superior performance. Similar to supervised research, the performance on the THUMOS14 benchmark has shown rapid improvement, while ActivityNet v1.2 and ActivityNet v1.3 present greater challenges and receive slower performance gains.

TABLE 7: Statistics about the size of existing temporal action localization datasets.

Dataset	Category No.	Video No.	Instance No.
THUMOS14 [192]	20	413	6,365
ActivityNet v1.2 [131]	100	9,682	10,733
FineAction [193]	106	17,000	103,000
ActivityNet v1.3 [131]	200	19,994	23,064
HACS [194]	200	50,000	140,000
FineGym v1.0 [195]	530	303	32,697
YouTube-8M [196]	1,000	46,000	237,000

"No." is the abbreviation for "Number".

4.4 Further discussions and promising directions

As discussed in the above sections, existing weakly temporal action localization research has achieved impressive progress. However, the community still requires continuous development in the following aspects: the size of the benchmark and the supervision label.

From small benchmark to large benchmark. Because a majority of temporal action localization algorithms are verified on THUMOS14 [192] and ActivityNet [131] benchmarks, current research is usually subject to two limitations: the risk of overfitting and the insufficient ability to extend. As shown in Table 7, the early benchmarks THUMOS14 and ActivityNet v1.3 contains 413 and 19,994 videos, respectively, while recently proposed benchmarks HACS [194] and YouTube-8M [196] contain 50,000 and 46,000 videos, respectively. In part because of overfitting, a classical method SSN [105] achieves 28.28% on ActivityNet v1.3, but the performance drops to 18.97% on HACS, although these two datasets share the same action categories. In addition, because existing algorithms are trained on fixed categories, the algorithm can accurately discover action instances belonging to categories in the benchmark but is incompetent when tackling newly emerged action categories. To alleviate this issue, Zhang et al. [197] have introduced zero-shot learning [198] to temporal action localization, but there exhibits huge performance gap with existing algorithms [17], [44], [52].

To summarize, we propose further research to benchmark the proposed methods on both large-scale datasets [194], [196], and datasets [193], [195] featuring fine-grained annotations. Regarding large-scale datasets, their utilization offers multiple advantages. Firstly, a large-scale benchmark provides an ample training data source that holds the potential to unlock the algorithm's full capabilities. Secondly, it poses new challenges due to the considerable computational burden and the diversity of instances within the dataset. As for fine-grained action analysis, the datasets FineAction [193] and FineGym v1.0 [195] have emerged as significant contributors to this research direction. In contrast to traditional coarse-level action annotations, fine-grained action analysis allows for atomic analyses, which are highly applicable in real-world scenarios. Moreover, it opens promising avenues for weakly-supervised learning in the context of fine-grained actions, as exemplified by the pioneering work of Li et al. [199].

Moreover, action localization algorithms should transcend conventional benchmarks, address novel categories, and excel in the challenging task of temporal action localization in unconstrained real-world settings. Recently, Chen et al. [200] integrated the weakly-supervised setting with open-set action localization, introducing a more formidable task. The recent advancements in

TABLE 8: Match rate of existing action localization methods on THUMOS14 dataset.

Method	Fully supervised methods			Weakly supervised methods		
	SSAD [33]	GTAD [16]	AFSD [44]	BaSNet [134]	DGAM [54]	UM [150]
Match Rate	71.58	85.18	92.68	56.29	52.64	62.47

TABLE 9: Analysis of existing action localization methods on THUMOS14 dataset.

Method	Fully supervised methods			Weakly supervised methods		
	SSAD [33]	GTAD [16]	AFSD [44]	BaSNet [134]	DGAM [54]	UM [150]
Raw method	48.00	43.03	55.48	28.69	28.11	33.81
Refining classification label	48.84 (+0.84)	43.36 (+0.33)	59.18 (+3.70)	28.90 (+0.21)	28.74 (+0.63)	33.98 (+0.17)
Refining confidence score	64.27 (+16.27)	82.83 (+39.80)	88.47 (+32.99)	56.57 (+27.88)	51.08 (+22.97)	61.28 (+27.47)

large foundational models [201], [202] have presented opportunities for significant progress in multimodal action localization tasks. These tasks can now leverage aligned vision-language features and factual knowledge encapsulated within the foundational models.

From fixed supervision to flexible supervision. Supervised temporal action localization necessitates accurately annotated action boundaries and classification labels. Conversely, weakly supervised algorithms remove precise annotations and learn directly from video-level classification labels. Ji et al. [203] alleviate the challenge of controlling annotation costs by proposing the semi-supervised action localization task. The primary hurdle in the semi-supervised setting lies in effectively leveraging limited annotations. To this end, several approaches have been explored: KFC [204] applies perturbations on video features, SSTAP [205] integrates self-supervised learning, Shi et al. [206] design three levels of supervision (fully-, weakly-, and unsupervised settings), and SPOT [207] introduces a parallel architecture for localization and classification. Apart from the semi-supervised setting, Kang et al. [208] developed the temporal action localization task in an online setting, constraining the algorithm to process streaming videos without accessing future frames. Subsequently, Kim et al. [209] improved the online setting by integrating the anchor mechanism.

More recently, Ma et al. [210] discovered that single-frame supervision yields performance gains comparable to weakly supervised methods but requires similar annotation costs. As a result, SF-Net [210] introduces the single-frame annotation to learn action patterns, which is developed by Yang et al. [136] via switching point annotations to backgrounds. Although current weakly supervised algorithms cannot achieve similar performance with their supervised counterparts, we think it is a promising direction to continuously develop the weakly supervised algorithm. With the rapid increase of video data on the internet, further works could strive to continuously obtain training videos from the internet and learn from a large amount of data. If so, the weakly supervised learning algorithm would persistently make improvements by itself, and such continual improvement may lead to a breakthrough.

In addition, the supervision signal should not be limited to instance- or video-level labels. For example, some works [211], [212] first employ trimmed videos from video recognition benchmarks to learn action patterns, then localize action instances in untrimmed videos. Furthermore, the exploration of multiple modalities within video data is essential. For instance, Lee et al. [213] introduced audio cues to enhance action localization methods. POLO [214] takes into account both appearance and motion modalities for action localization. Additionally, the localization of action instances based on language descriptions [215], often referred to as temporal grounding [216], [217], has been extensively studied. Moreover, Bao et al. [218] propose the open-set temporal action localization task. In this task, the algorithm is trained with fixed action categories but is expected to accurately detect unknown actions when handling videos containing previously unseen categories. Given the abundance of supervision signals and multimodal video data, we recommend that future research focuses on building a flexible learning framework capable of exploring various supervisions and effectively leveraging the complementarity among multiple modalities. -

5 Future Directions

5.1 Improving confidence score estimation

As extensively reviewed above, the temporal action localization task has been extensively investigated through a series of seminal works [1], [2], [13], [14], [15], [16], [17], [19], [20], [21], [151]. Overall, the process of localizing an action instance involves three key subtasks: precisely identifying action boundaries, accurately classifying the action category, and appropriately estimating the confidence score. While past research has primarily focused on the first two subtasks, the significance of the confidence score has not received sufficient attention, leading to a performance bottleneck in the existing approaches.

To identify the performance bottleneck of current methods, a natural approach is to employ the diagnosing tool DETAD [219] and conduct an analysis of the localization results. DETAD [219] categorizes false positive predictions into five categories: Double Detection Error, Wrong Label Error, Localization Error, Confusion Error, and Background Error. However, the last four categories only consider the boundary and classification quality, neglecting the influence of the confidence score. Furthermore, the Double Detection Error only indicates the presence of another true positive prediction with a higher confidence score, without analyzing confidence scores among multiple predictions. As our objective is to analyze the impact of the confidence score, a factor that has been overlooked for a considerable period, we conduct experiments under an oracle setting and analyze both supervised methods [16], [33], [44], and weakly supervised methods [54], [134], [150], as illustrated in Table 9.

Given the localization results of each method on the THUMOS14 dataset, we first evaluate the quality of action proposals.

Specifically, for each ground truth instance, we measure the Intersection over Union (IoU) between the ground truth instance and any predicted instance. If the predicted instance has a high IoU value (i.e., IoU $>$ 0.5) with the ground truth instance, we consider this ground truth instance as correctly matched. As demonstrated in Table 8, existing methods exhibit a high match rate, with AFSD [44] achieving a ground truth match rate of 92.68%. Next, as shown in Table 9, we refine the classification label by assigning each prediction the label of the closest ground truth instance. However, this refinement only results in limited performance gains, typically less than 1.0%, which suggests that the classification task is not the primary performance bottleneck. On the other hand, if we refine the confidence score based on the IoU between a predicted instance and its closest ground truth, we observe substantial performance gains. For instance, the supervised method AFSD [44] achieves a gain of 32.99% and an mAP of 88.47%. Even the weakly supervised method UM [150] achieves an mAP of 61.28% with the proposed refinement. This notable improvement indicates that the main performance bottleneck lies in the accurate estimation of the confidence score. In summary, our comprehensive analysis demonstrates that the confidence score estimation significantly influences the overall performance of the temporal action localization methods, and addressing this aspect can lead to considerable performance gains.

5.2 Pursuing high-quality feature representation

Discriminative features play a crucial role in assisting action localization algorithms by distinguishing action frames from contextual backgrounds and enabling differentiation among multiple action categories. Moreover, representative features help reduce intra-class diversity and increase inter-class differences, further enhancing the algorithm's performance. However, a significant number of existing methods rely on off-the-shelf video recognition backbones [69], [70], [71] for feature extraction from untrimmed videos. Unfortunately, these backbones are not flexible enough due to their inherent task bias. To pursue the extraction of high-quality features, we propose exploring unsupervised visual representation learning [220], [221] as a promising direction, particularly in the context of video-related tasks [222], [223], [224]. By leveraging unsupervised learning, these specialized backbones can alleviate the task bias and produce features that encapsulate both intra-class similarity and inter-class separability. This approach holds great potential for improving the overall performance of action localization algorithms by providing more discriminative and representative features.

5.3 Preserving privacy

In the data-driven learning paradigm, the performance of deep learning algorithms is often directly influenced by the quantity of available training data. Consequently, a larger dataset with more action instances can significantly enhance the temporal action localization algorithm's performance. However, the abundance of data also raises concerns about privacy and potential misuse of personal information [225]. Without appropriate privacy preservation measures, the action localization system may inadvertently expose sensitive information for illicit purposes. To address this issue and effectively preserve privacy, two promising strategies have emerged: secure multi-party learning [226], [227] and federated learning [228]. Secure multi-party learning utilizes a peer-to-peer pipeline, where multiple parties collaboratively train the model without sharing their individual data directly. On the other hand, federated learning adopts a client-server pipeline, where clients with private data locally train their models and only share model updates with the central server. Both approaches are designed to ensure data privacy while allowing the algorithm to leverage information from multiple sources. It is important to note that privacy preservation should be considered not only during the training phase but also during the inference phase. Proper precautions and secure mechanisms must be in place to guarantee that personal information remains protected throughout the entire action localization process.

5.4 Measuring video information

In theory, a fundamental challenge in accurately localizing action instances from untrimmed videos lies in the precise extraction of meaningful information from the video content. If we can extend traditional information theory [229] to the video understanding domain and establish effective rules for measuring video information, it could lead to significant breakthroughs in temporal action localization research. For instance, a series of video frames containing the desired information could form a coherent action instance. Moreover, with such a framework, it might be possible to quantitatively assess the impact of each video processing step on action localization results, thereby providing valuable insights for advancing research in intelligent video understanding.

6 Conclusion

This paper presents a comprehensive survey of temporal action localization methods in the deep learning era. For supervised action localization, we thoroughly review the anchor mechanism and provide a detailed discussion of the strengths and weaknesses of one-stage, two-stage, and anchor-free pipelines. Furthermore, we summarize existing works in the classification mechanism, encompassing both the frame classification and proposal classification pipelines, offering an in-depth and novel overview of supervised action localization methods. In the domain of weakly supervised temporal action localization, we categorize existing works into pre-classification and post-classification pipelines, carefully examining their distinct characteristics. Additionally, we conduct a comprehensive analysis of the classifier, feature, and attention weight, leading to a novel categorization of weakly supervised methods. The taxonomy we propose enhances the understanding of temporal action localization research and facilitates comparisons between different approaches.

Moreover, this survey uncovers the performance bottleneck related to confidence estimation. Looking ahead, we suggest several promising research directions. One such direction involves the precise quantification of information within a video to aid the development of intelligent video understanding. Additionally, we emphasize the significance of preserving privacy in action localization systems, proposing secure multi-party learning and federated learning as viable strategies. In conclusion, this survey contributes to the field by offering a comprehensive overview of temporal action localization methods and their strengths and limitations. The suggested future research directions serve as valuable insights for researchers in the community.

References

[1] Z. Shou, D. Wang, and S.-F. Chang, "Temporal action localization in untrimmed videos via multi-stage cnns," in *CVPR*, 2016, pp. 1049–1058. 1, 2, 3, 4, 5, 6, 7, 8, 13

[2] L. Wang, Y. Xiong, D. Lin, and L. Van Gool, "Untrimmednets for weakly supervised action recognition and detection," in *CVPR*, 2017, pp. 4325–4334. 1, 2, 3, 8, 9, 10, 11, 12, 13

[3] A. Tejero-de Pablos, Y. Nakashima, T. Sato, N. Yokoya, M. Linna, and E. Rahtu, "Summarization of user-generated sports video by using deep action recognition features," *TMM*, vol. 20, no. 8, pp. 2000–2011, 2018. 1

[4] C. Loukas, "Video content analysis of surgical procedures," *Surgical endoscopy*, vol. 32, no. 2, pp. 553–568, 2018. 1

[5] Y. Jiao, Z. Li, S. Huang, X. Yang, B. Liu, and T. Zhang, "Three-dimensional attention-based deep ranking model for video highlight detection," *TMM*, vol. 20, no. 10, pp. 2693–2705, 2018. 1

[6] B. Zhao, X. Li, and X. Lu, "Property-constrained dual learning for video summarization," *TNNLS*, vol. 31, no. 10, pp. 3989–4000, 2019. 1

[7] X. Li and B. Zhao, "Video distillation," *SCIENTIA SINICA Informationis*, vol. 51, pp. 695–734, 2021. 1

[8] J. Lee and S. Abu-El-Haija, "Large-scale content-only video recommendation," in *ICCVW*, 2017, pp. 987–995. 1

[9] X. Li and Z. Zhao, "Pixel level semantic understanding: from classification to regression," *SCIENTIA SINICA Informationis*, vol. 51, pp. 521–564, 2021. 1

[10] K.-E. Ko and K.-B. Sim, "Deep convolutional framework for abnormal behavior detection in a smart surveillance system," *Engineering Applications of Artificial Intelligence*, vol. 67, pp. 226–234, 2018. 1

[11] K. Yun, Y. Kwon, S. Oh, J. Moon, and J. Park, "Vision-based garbage dumping action detection for real-world surveillance platform," *ETRI Journal*, vol. 41, no. 4, pp. 494–505, 2019. 1

[12] A. Vignolo, N. Noceti, F. Rea, A. Sciutti, F. Odone, and G. Sandini, "Detecting biological motion for human–robot interaction: A link between perception and action," *Frontiers in Robotics and AI*, vol. 4, p. 14, 2017. 1

[13] Z. Y. Jiyang Gao and R. Nevatia, "Cascaded boundary regression for temporal action detection," in *BMVC*, 2017, pp. 52.1–52.11. 2, 4, 5, 6, 7, 13

[14] Y.-W. Chao, S. Vijayanarasimhan, B. Seybold, D. A. Ross, J. Deng, and R. Sukthankar, "Rethinking the faster r-cnn architecture for temporal action localization," in *CVPR*, 2018, pp. 1130–1139. 2, 3, 4, 5, 6, 7, 8, 13

[15] R. Zeng, W. Huang, M. Tan, Y. Rong, P. Zhao, J. Huang, and C. Gan, "Graph convolutional networks for temporal action localization," in *ICCV*, 2019, pp. 7094–7103. 2, 3, 4, 5, 6, 7, 13

[16] M. Xu, C. Zhao, D. S. Rojas, A. Thabet, and B. Ghanem, "G-tad: Sub-graph localization for temporal action detection," in *CVPR*, 2020, pp. 10 156–10 165. 2, 3, 4, 6, 7, 13

[17] X. Liu, Y. Hu, S. Bai, F. Ding, X. Bai, and P. H. Torr, "Multi-shot temporal event localization: a benchmark," in *CVPR*, 2021, pp. 12 596–12 606. 2, 4, 5, 6, 7, 12, 13

[18] C.-L. Zhang, J. Wu, and Y. Li, "Actionformer: Localizing moments of actions with transformers," in *ECCV*. Springer, 2022, pp. 492–510. 2, 4, 5, 6, 7

[19] S. Paul, S. Roy, and A. K. Roy-Chowdhury, "W-talc: Weakly-supervised temporal activity localization and classification," in *ECCV*, 2018, pp. 563–579. 2, 8, 9, 11, 13

[20] P. X. Nguyen, D. Ramanan, and C. C. Fowlkes, "Weakly-supervised action localization with background modeling," in *ICCV*, 2019, pp. 5502–5511. 2, 3, 8, 9, 10, 11, 12, 13

[21] Z. Luo, D. Guillory, B. Shi, W. Ke, F. Wan, T. Darrell, and H. Xu, "Weakly-supervised action localization with expectation-maximization multi-instance learning," in *ECCV*, 2020, pp. 729–745. 2, 9, 10, 11, 13

[22] F.-T. Hong, J.-C. Feng, D. Xu, Y. Shan, and W.-S. Zheng, "Cross-modal consensus network for weakly supervised temporal action localization," in *ACMMM*, 2021, pp. 1591–1599. 2, 10, 11

[23] L. Huang, Y. Huang, W. Ouyang, and L. Wang, "Modeling sub-actions for weakly supervised temporal action localization," *TIP*, vol. 30, pp. 5154–5167, 2021. 2, 9, 10, 12

[24] D. Shi, Y. Zhong, Q. Cao, L. Ma, J. Li, and D. Tao, "Tridet: Temporal action detection with relative boundary modeling," in *CVPR*, 2023, pp. 18 857–18 866. 1, 2, 4, 7

[25] M. N. Rizve, G. Mittal, Y. Yu, M. Hall, S. Sajeev, M. Shah, and M. Chen, "Pivotal: Prior-driven supervision for weakly-supervised temporal action localization," in *CVPR*, 2023, pp. 22 992–23 002. 2, 8, 10, 11

[26] H. Xia and Y. Zhan, "A survey on temporal action localization," *IEEE Access*, vol. 8, pp. 70 477–70 487, 2020. 2, 3

[27] E. Vahdani and Y. Tian, "Deep learning-based action detection in untrimmed videos: A survey," *TPAMI*, 2022. 2, 3

[28] A. Baraka and M. H. Mohd Noor, "Weakly-supervised temporal action localization: a survey," *NCA*, pp. 1–21, 2022. 3

[29] R. Poppe, "A survey on vision-based human action recognition," *Image and vision computing*, vol. 28, no. 6, pp. 976–990, 2010. 2

[30] S. Herath, M. Harandi, and F. Porikli, "Going deeper into action recognition: A survey," *Image and vision computing*, vol. 60, pp. 4–21, 2017. 2

[31] Y. Kong and Y. Fu, "Human action recognition and prediction: A survey," *IJCV*, vol. 130, no. 5, pp. 1366–1401, 2022. 2

[32] S. Yeung, O. Russakovsky, G. Mori, and L. Fei-Fei, "End-to-end learning of action detection from frame glimpses in videos," in *CVPR*, 2016, pp. 2678–2687. 4, 5, 7

[33] T. Lin, X. Zhao, and Z. Shou, "Single shot temporal action detection," in *ACMMM*, 2017, pp. 988–996. 3, 4, 5, 6, 7, 8, 13

[34] Z. Shou, J. Chan, A. Zareian, K. Miyazawa, and S.-F. Chang, "Cdc: Convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos," in *CVPR*, 2017, pp. 5734–5743. 4, 5, 6, 7, 8

[35] S. Buch, V. Escorcia, B. Ghanem, L. Fei-Fei, and J. C. Niebles, "End-to-end, single-stream temporal action detection in untrimmed videos," in *BMVC*, vol. 2, 2017, p. 7. 4, 5, 6, 7

[36] H. Xu, A. Das, and K. Saenko, "R-c3d: Region convolutional 3d network for temporal activity detection," in *ICCV*, 2017, pp. 5783–5792. 3, 4, 5, 6, 7, 8

[37] K. Yang, P. Qiao, D. Li, S. Lv, and Y. Dou, "Exploring temporal preservation networks for precise temporal action localization," in *AAAI*, vol. 32, no. 1, 2018. 4, 5, 6

[38] T. Lin, X. Zhao, H. Su, C. Wang, and M. Yang, "Bsn: Boundary sensitive network for temporal action proposal generation," in *ECCV*, 2018, pp. 3–19. 4, 5, 6, 7

[39] F. Long, T. Yao, Z. Qiu, X. Tian, J. Luo, and T. Mei, "Gaussian temporal awareness networks for action localization," in *CVPR*, 2019, pp. 344–353. 3, 4, 5, 6, 7

[40] A. Piergiovanni and M. S. Ryoo, "Temporal gaussian mixture layer for videos," *Proceedings of Machine Learning Research*, vol. 97, 2019. 4, 5, 6

[41] P. Zhao, L. Xie, C. Ju, Y. Zhang, Y. Wang, and Q. Tian, "Bottom-up temporal action localization with mutual regularization," in *ECCV*, 2020, pp. 539–555. 4, 5, 6, 7, 8

[42] T. Lin, X. Liu, X. Li, E. Ding, and S. Wen, "Bmn: Boundary-matching network for temporal action proposal generation," in *ICCV*, 2019, pp. 3889–3898. 3, 4, 5, 6, 7

[43] J. Li, X. Liu, Z. Zong, W. Zhao, M. Zhang, and J. Song, "Graph attention based proposal 3d convnets for action detection," in *AAAI*, vol. 34, no. 04, 2020, pp. 4626–4633. 4, 5, 6

[44] C. Lin, C. Xu, D. Luo, Y. Wang, Y. Tai, C. Wang, J. Li, F. Huang, and Y. Fu, "Learning salient boundary feature for anchor-free temporal action localization," in *CVPR*, 2021, pp. 3320–3329. 3, 4, 5, 6, 7, 12, 13, 14

[45] R. Zeng, W. Huang, M. Tan, Y. Rong, P. Zhao, J. Huang, and C. Gan, "Graph convolutional module for temporal action localization in videos," *TPAMI*, 2021. 4, 5, 6, 7

[46] D. Sridhar, N. Quader, S. Muralidharan, Y. Li, P. Dai, and J. Lu, "Class semantics-based attention for action detection," in *ICCV*, 2021, pp. 13 739–13 748. 4, 5, 6

[47] F. Cheng and G. Bertasius, "Tallformer: Temporal action localization with a long-memory transformer," in *ECCV*. Springer, 2022, pp. 503–521. 4, 5, 6, 7

[48] L. Yang, J. Han, T. Zhao, N. Liu, and D. Zhang, "Structured attention composition for temporal action localization," *TIP*, 2022. 4

[49] Z. Zhu, L. Wang, W. Tang, N. Zheng, and G. Hua, "Contextloc++: A unified context model for temporal action localization," *TPAMI*, 2023. 4, 5, 6, 7

[50] Q. Liu and Z. Wang, "Progressive boundary refinement network for temporal action detection," in *AAAI*, vol. 34, no. 07, 2020, pp. 11 612–11 619. 3, 5, 7

[51] Y. Liu, L. Ma, Y. Zhang, W. Liu, and S.-F. Chang, "Multi-granularity generator for temporal action proposal," in *CVPR*, 2019, pp. 3604–3613. 3, 5, 6

[52] Z. Qing, H. Su, W. Gan, D. Wang, W. Wu, X. Wang, Y. Qiao, J. Yan, C. Gao, and N. Sang, "Temporal context aggregation network for temporal action proposal refinement," in *CVPR*, 2021, pp. 485–494. 3, 5, 6, 7, 12

[53] S. Narayan, H. Cholakkal, F. S. Khan, and L. Shao, "3c-net: Category count and center loss for weakly-supervised action localization," in *ICCV*, 2019, pp. 8679–8687. 3, 8, 9, 10, 11, 12

[54] B. Shi, Q. Dai, Y. Mu, and J. Wang, "Weakly-supervised action localization by generative attention modeling," in *CVPR*, 2020, pp. 1009–1019. 3, 8, 10, 11, 12, 13

[55] Z. Liu, L. Wang, Q. Zhang, W. Tang, J. Yuan, N. Zheng, and G. Hua, "Acsnet: Action-context separation network for weakly supervised temporal action localization," in *AAAI*, vol. 35, no. 3, 2021, pp. 2233–2241. 3, 8, 10, 11, 12

[56] Z. Liu, L. Wang, J. Yuan, N. Zheng, and G. Hua, "Weakly supervised temporal action localization through learning explicit subspaces for action and context," in *AAAI*, 2021. 3, 10, 11, 12

[57] K. K. Singh and Y. J. Lee, "Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization," in *ICCV*, 2017, pp. 3544–3553. 3, 9, 10, 11, 12

[58] D. Liu, T. Jiang, and Y. Wang, "Completeness modeling and context separation for weakly supervised temporal action localization," in *CVPR*, 2019, pp. 1298–1307. 3, 8, 9, 10, 11, 12

[59] K. Min and J. J. Corso, "Adversarial background-aware loss for weakly-supervised temporal activity localization," in *ECCV*, 2020, pp. 283–299. 3, 9, 10, 11, 12

[60] T. Zhao, J. Han, L. Yang, B. Wang, and D. Zhang, "Soda: Weakly supervised temporal action localization based on astute background response and self-distillation learning," *IJCV*, pp. 1–25, 2021. 3

[61] L. Zelnik-Manor and M. Irani, "Event-based analysis of video," in *CVPR*, vol. 2, 2001, pp. II–II. 3

[62] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *TPAMI*, vol. 29, no. 12, pp. 2247–2253, 2007. 3

[63] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *CVPR*, 2008, pp. 1–8. 3

[64] O. Duchenne, I. Laptev, J. Sivic, F. Bach, and J. Ponce, "Automatic annotation of human actions in video," in *ICCV*, 2009, pp. 1491–1498. 3

[65] W. Brendel and S. Todorovic, "Learning spatiotemporal graphs of human activities," in *ICCV*, 2011, pp. 778–785. 3

[66] K. Tang, L. Fei-Fei, and D. Koller, "Learning latent temporal structure for complex event detection," in *CVPR*, 2012, pp. 1250–1257. 3

[67] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *NeurIPS*, vol. 25, pp. 1097–1105, 2012. 3

[68] K. Simonyan and A. Zisserman, "very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015. 3, 7

[69] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *ICCV*, 2015, pp. 4489–4497. 3, 7, 8, 14

[70] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *CVPR*, 2017, pp. 6299–6308. 3, 7, 8, 12, 14

[71] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *ICCV*, 2019, pp. 6202–6211. 3, 8, 14

[72] Y. Huang, Q. Dai, and Y. Lu, "Decoupling localization and classification in single shot temporal action detection," in *ICME*, 2019, pp. 1288–1293. 3, 4, 5

[73] C. Wang, H. Cai, Y. Zou, and Y. Xiong, "Rgb stream is enough for temporal action detection," *arXiv preprint arXiv:2107.04362*, 2021. 3, 4, 5, 6, 7, 8

[74] H. Xu, A. Das, and K. Saenko, "Two-stream region convolutional 3d network for temporal activity detection," *TPAMI*, vol. 41, no. 10, pp. 2319–2332, 2019. 3, 5

[75] L. Li, T. Kong, F. Sun, and H. Liu, "Deep point-wise prediction for action temporal proposal," in *International conference on neural information processing*, 2019, pp. 475–487. 3, 5, 6

[76] Y. Tang, C. Niu, M. Dong, S. Ren, and J. Liang, "Afo-tad: Anchor-free one-stage detector for temporal action detection," *arXiv preprint arXiv:1910.08250*, 2019. 3, 5, 6

[77] L. Yang, H. Peng, D. Zhang, J. Fu, and J. Han, "Revisiting anchor mechanisms for temporal action localization," *TIP*, vol. 29, pp. 8535–8548, 2020. 3, 5, 6, 7

[78] Z. Shou, H. Gao, L. Zhang, K. Miyazawa, and S.-F. Chang, "Autoloc: Weakly-supervised temporal action localization in untrimmed videos," in *ECCV*, 2018, pp. 154–171. 3, 8, 9, 10, 11

[79] C. Zhang, Y. Xu, Z. Cheng, Y. Niu, S. Pu, F. Wu, and F. Zou, "Adversarial seeded sequence growing for weakly-supervised temporal action localization," in *ACMMM*, 2019, pp. 738–746. 3, 9, 10, 11

[80] C. Zhang, M. Cao, D. Yang, J. Chen, and Y. Zou, "Cola: Weakly-supervised temporal action localization with snippet contrastive learning," in *CVPR*, 2021, pp. 16 010–16 019. 3, 8, 10, 11

[81] P. Nguyen, T. Liu, G. Prasad, and B. Han, "Weakly supervised action localization by sparse temporal pooling network," in *CVPR*, 2018, pp. 6752–6761. 3, 8, 10, 11

[82] W. Yang, T. Zhang, Z. Mao, Y. Zhanga, Q. Tian, and F. Wu, "Multi-scale structure-aware network for weakly supervised temporal action detection," *TIP*, 2021. 3, 10, 11

[83] Y. A. Farha and J. Gall, "Ms-tcn: Multi-stage temporal convolutional network for action segmentation," in *CVPR*, 2019, pp. 3575–3584. 3

[84] Y. Li, W. Lin, J. See, N. Xu, S. Xu, K. Yan, and C. Yang, "Cfad: Coarse-to-fine action detector for spatiotemporal action localization," in *ECCV*, 2020, pp. 510–527. 3

[85] J. Gao, Z. Yang, K. Chen, C. Sun, and R. Nevatia, "Turn tap: Temporal unit regression network for temporal action proposals," in *ICCV*, 2017, pp. 3628–3636. 4, 5, 7

[86] M. A. Rahman and R. Laganière, "Mid-level fusion for end-to-end temporal activity detection in untrimmed video." in *BMVC*, 2020. 4, 5, 6

[87] X. Li, T. Lin, X. Liu, W. Zuo, C. Li, X. Long, D. He, F. Li, S. Wen, and C. Gan, "Deep concept-wise temporal convolutional networks for action localization," in *ACMMM*, 2020, pp. 4004–4012. 5, 6, 7

[88] H. Yang, W. Wu, L. Wang, S. Jin, B. Xia, H. Yao, and H. Huang, "Temporal action proposal generation with background constraint," in *AAAI*, 2022. 5, 6

[89] G. Chen, C. Zhang, and Y. Zou, "Afnet: Temporal locality-aware network with dual structure for accurate and fast action detection," *TMM*, 2020. 5, 7

[90] Z. Zhu, W. Tang, L. Wang, N. Zheng, and G. Hua, "Enriching local and global contexts for temporal action localization," in *ICCV*, 2021, pp. 13 516–13 525. 5, 6, 7

[91] H. Kang, H. Kim, J. An, M. Cho, and S. J. Kim, "Soft-landing strategy for alleviating the task discrepancy problem in temporal action localization tasks," in *CVPR*, 2023, pp. 6514–6523. 5, 6, 7

[92] Q. Wang, Y. Zhang, Y. Zheng, and P. Pan, "Rcl: Recurrent continuous localization for temporal action detection," in *CVPR*, 2022, pp. 13 566–13 575. 5, 6, 7

[93] R. Ning, C. Zhang, and Y. Zou, "Srf-net: Selective receptive field network for anchor-free temporal action detection," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 2460–2464. 5, 6

[94] D. Shi, Y. Zhong, Q. Cao, J. Zhang, L. Ma, J. Li, and D. Tao, "React: Temporal action detection with relational queries," in *ECCV*. Springer, 2022, pp. 105–121. 5, 6, 7

[95] X. Liu, Q. Wang, Y. Hu, X. Tang, S. Zhang, S. Bai, and X. Bai, "End-to-end temporal action detection with transformer," *TIP*, vol. 31, pp. 5427–5441, 2022. 5, 6, 7

[96] C. Zhao, S. Liu, K. Mangalam, and B. Ghanem, "Re2tal: Rewiring pretrained video backbones for reversible temporal action localization," in *CVPR*, 2023, pp. 10 637–10 647. 5, 6, 7, 8

[97] S. Ma, L. Sigal, and S. Sclaroff, "Learning activity progression in lstms for activity detection and early detection," in *CVPR*, 2016, pp. 1942–1950. 5, 6, 7

[98] J. Gao, K. Chen, and R. Nevatia, "Ctap: Complementary temporal action proposal generation," in *ECCV*, 2018, pp. 68–83. 5, 6, 7, 8

[99] H. Qiu, Y. Zheng, H. Ye, Y. Lu, F. Wang, and L. He, "Precise temporal action localization by evolving temporal proposals," in *AAAI*, 2018, pp. 388–396. 5, 6

[100] C. Zhao, A. K. Thabet, and B. Ghanem, "Video self-stitching graph network for temporal action localization," in *ICCV*, 2021, pp. 13 658–13 667. 5

[101] S. Chang, P. Wang, F. Wang, H. Li, and J. Feng, "Augmented transformer with adaptive graph for temporal action proposal generation," *arXiv preprint arXiv:2103.16024*, 2021. 5, 6

[102] C. Sun, H. Song, X. Wu, Y. Jia, and J. Luo, "Exploiting informative video segments for temporal action localization," *TMM*, 2021. 5, 6

[103] K. Xia, L. Wang, S. Zhou, N. Zheng, and W. Tang, "Learning to refactor action and co-occurrence features for temporal action localization," in *CVPR*, 2022, pp. 13 884–13 893. 5, 6, 7

[104] G. Chen, Y.-D. Zheng, L. Wang, and T. Lu, "Dcan: improving temporal action detection via dual context aggregation," in *AAAI*, vol. 36, no. 1, 2022, pp. 248–257. 5, 6, 7

[105] Y. Zhao, Y. Xiong, L. Wang, Z. Wu, X. Tang, and D. Lin, "Temporal action detection with structured segment networks," in *ICCV*, 2017, pp. 2914–2923. 5, 6, 7, 12

[106] S. Buch, V. Escorcia, C. Shen, B. Ghanem, and J. Carlos Niebles, "Sst: Single-stream temporal action proposals," in *CVPR*, 2017, pp. 2911–2920. 5, 6, 7

[107] P. Chen, C. Gan, G. Shen, W. Huang, R. Zeng, and M. Tan, "Relation attention for temporal action localization," *TMM*, vol. 22, no. 10, pp. 2723–2733, 2019. 5, 6

[108] Y. Zhou, R. Wang, H. Li, and S. Y. Kung, "Temporal action localization using long short-term dependency," *TMM*, 2020. 5, 6, 7

[109] H. Su, W. Gan, W. Wu, Y. Qiao, and J. Yan, "Bsn++: Complementary boundary regressor with scale-balanced relation modeling for temporal action proposal generation," in *AAAI*, vol. 35, no. 3, 2021, pp. 2602–2610. 5, 6, 7

[110] Y. Bai, Y. Wang, Y. Tong, Y. Yang, Q. Liu, and J. Liu, "Boundary content graph neural network for temporal action proposal generation," in *ECCV*, 2020, pp. 121–137. 5, 6

[111] J. Tan, J. Tang, L. Wang, and G. Wu, "Relaxed transformer decoders for direct action proposal generation," in *ICCV*, 2021, pp. 13 526–13 535. 6

[112] M. Nawhal and G. Mori, "Activity graph transformer for temporal action localization," *arXiv preprint arXiv:2101.08540*, 2021. 6

[113] X. Liu, S. Bai, and X. Bai, "An empirical study of end-to-end temporal action detection," in *CVPR*, 2022, pp. 20 010–20 019. 6

[114] W. Wang, Y. Wu, H. Liu, S. Wang, and J. Cheng, "Temporal action detection by joint identification-verification," in *ICPR*, 2018, pp. 2026–2031. 6

[115] W. Zhang, B. Wang, S. Ma, Y. Zhang, and Y. Zhao, "I2net: Mining intra-video and inter-video attention for temporal action localization," *Neurocomputing*, vol. 444, pp. 16–29, 2021. 6

[116] H. Song, X. Wu, B. Zhu, Y. Wu, M. Chen, and Y. Jia, "Temporal action localization in untrimmed videos using action pattern trees," *TMM*, vol. 21, no. 3, pp. 717–730, 2018. 6

[117] M. Xu, J.-M. Perez-Rua, X. Zhu, B. Ghanem, and B. Martinez, "Low-fidelity end-to-end video encoder pre-training for temporal action localization," in *NeurIPS*, 2021. 6

[118] F. C. Heilbron, J. C. Niebles, and B. Ghanem, "Fast temporal activity proposals for efficient detection of human actions in untrimmed videos," in *CVPR*, 2016, pp. 1914–1923. 6, 7, 8

[119] R. Su, D. Xu, L. Sheng, and W. Ouyang, "Pcg-tal: Progressive cross-granularity cooperation for temporal action localization," *TIP*, vol. 30, pp. 2103–2113, 2020. 6, 7

[120] Z. Yuan, J. C. Stroud, T. Lu, and J. Deng, "Temporal action localization by structured maximal sums," in *CVPR*, 2017, pp. 3684–3692. 6

[121] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," in *ICCV*, 2019, pp. 9627–9636. 5

[122] V. Escorcia, F. C. Heilbron, J. C. Niebles, and B. Ghanem, "Daps: Deep action proposals for action understanding," in *ECCV*, 2016, pp. 768–784. 6, 7

[123] Y. Xiong, Y. Zhao, L. Wang, D. Lin, and X. Tang, "A pursuit of temporal accuracy in general activity detection," *arXiv preprint arXiv:1703.02716*, 2017. 6, 7

[124] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks for action recognition in videos," *TPAMI*, vol. 41, no. 11, pp. 2740–2755, 2018. 7

[125] S. Nag, X. Zhu, Y.-Z. Song, and T. Xiang, "Proposal-free temporal action detection via global segmentation mask learning," in *ECCV*. Springer, 2022, pp. 645–662. 7

[126] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997. 6

[127] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," in *EMNLP*, 2014, pp. 1724–1734. 6

[128] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *ECCV*, 2016, pp. 20–36. 7, 8

[129] X. Lu, W. Wang, J. Shen, Y.-W. Tai, D. J. Crandall, and S. C. Hoi, "Learning video object segmentation from unlabeled videos," in *CVPR*, 2020, pp. 8960–8970. 7

[130] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778. 7

[131] B. G. Fabian Caba Heilbron, Victor Escorcia and J. C. Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *CVPR*, 2015, pp. 961–970. 7, 12

[132] Y. Yuan, Y. Lyu, X. Shen, I. W. Tsang, and D.-Y. Yeung, "Marginalized average attentional network for weakly-supervised learning," in *ICLR*, 2019. 8, 10, 11, 12

[133] Y. Zhai, L. Wang, W. Tang, Q. Zhang, J. Yuan, and G. Hua, "Two-stream consensus network for weakly-supervised temporal action localization," in *ECCV*, 2020, pp. 37–54. 8, 10, 11, 12

[134] P. Lee, Y. Uh, and H. Byun, "Background suppression network for weakly-supervised temporal action localization," in *AAAI*, vol. 34, no. 07, 2020, pp. 11 320–11 327. 8, 9, 10, 11, 13

[135] W. Luo, T. Zhang, W. Yang, J. Liu, T. Mei, F. Wu, and Y. Zhang, "Action unit memory network for weakly supervised temporal action localization," in *CVPR*, 2021, pp. 9969–9979. 8, 9, 10, 11, 12

[136] L. Yang, J. Han, T. Zhao, T. Lin, D. Zhang, and J. Chen, "Background-click supervision for temporal action localization," *TPAMI*, 2021. 8, 13

[137] T. Zhao, J. Han, L. Yang, and D. Zhang, "Equivalent classification mapping for weakly supervised temporal action localization," *TPAMI*, 2022. 8, 11

[138] B. He, X. Yang, L. Kang, Z. Cheng, X. Zhou, and A. Shrivastava, "Asm-loc: Action-aware segment modeling for weakly-supervised temporal action localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 925–13 935. 8, 10, 11

[139] M. Chen, J. Gao, S. Yang, and C. Xu, "Dual-evidential learning for weakly-supervised temporal action localization," in *ECCV*. Springer, 2022, pp. 192–208. 8, 10, 11

[140] J. Fu, J. Gao, and C. Xu, "Semantic and temporal contextual correlation learning for weakly-supervised temporal action localization," *TPAMI*, 2023. 8, 10, 11

[141] H. Alwassel, S. Giancola, and B. Ghanem, "Tsp: Temporally-sensitive pretraining of video encoders for localization tasks," in *ICCV*, 2021, pp. 3173–3183. 7

[142] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *NeurIPS*, 2014, pp. 568–576. 8

[143] D. Damen, H. Doughty, G. M. Farinella, A. Furnari, E. Kazakos, J. Ma, D. Moltisanti, J. Munro, T. Perrett, W. Price *et al.*, "Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100," *IJCV*, pp. 1–23, 2021. 8

[144] R. De Geest, E. Gavves, A. Ghodrati, Z. Li, C. Snoek, and T. Tuytelaars, "Online action detection," in *ECCV*, 2016, pp. 269–284. 8

[145] X. Wang, S. Zhang, Z. Qing, Y. Shao, Z. Zuo, C. Gao, and N. Sang, "Oadtr: Online action detection with transformers," in *ICCV*, 2021. 8

[146] J.-X. Zhong, N. Li, W. Kong, T. Zhang, T. H. Li, and G. Li, "Step-by-step erasion, one-by-one collection: a weakly supervised temporal action detector," in *ACMMM*, 2018, pp. 35–44. 9, 10

[147] Z. Liu, L. Wang, Q. Zhang, Z. Gao, Z. Niu, N. Zheng, and G. Hua, "Weakly supervised temporal action localization through contrast based evaluation networks," in *ICCV*, 2019, pp. 3899–3908. 9, 10, 11

[148] M. Moniruzzaman, Z. Yin, Z. He, R. Qin, and M. C. Leu, "Action completeness modeling with background aware networks for weakly-supervised temporal action localization," in *ACMMM*, 2020, pp. 2166–2174. 9, 10, 12

[149] A. Islam, C. Long, and R. Radke, "A hybrid attention mechanism for weakly-supervised temporal action localization," in *AAAI*, vol. 35, no. 2, 2021, pp. 1637–1645. 9, 10, 11, 12

[150] P. Lee, J. Wang, Y. Lu, and H. Byun, "Weakly-supervised temporal action localization by uncertainty modeling," in *AAAI*, vol. 35, no. 3, 2021, pp. 1854–1862. 9, 10, 11, 12, 13, 14

[151] Y. Liu, J. Chen, Z. Chen, B. Deng, J. Huang, and H. Zhang, "The blessings of unlabeled background in untrimmed videos," in *CVPR*, 2021, pp. 6176–6185. 9, 10, 11, 13

[152] J. Ma, S. K. Gorti, M. Volkovs, and G. Yu, "Weakly supervised action selection learning in video," in *CVPR*, 2021, pp. 7587–7596. 10, 11

[153] Y. Ji, X. Jia, H. Lu, and X. Ruan, "Weakly-supervised temporal action localization via cross-stream collaborative learning," in *ACMMM*, 2021, pp. 853–861. 10

[154] L. Huang, L. Wang, and H. Li, "Foreground-action consistency network for weakly supervised temporal action localization," in *ICCV*, 2021, pp. 8002–8011. 10

[155] S. Narayan, H. Cholakkal, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao, "D2-net: Weakly-supervised action localization via discriminative embeddings and denoised activations," in *ICCV*, 2021, pp. 13 608–13 617. 10, 11

[156] L. Huang, Y. Huang, W. Ouyang, and L. Wang, "Two-branch relational prototypical network for weakly supervised temporal action localization," *TPAMI*, 2021. 10, 11

[157] M. Cao, C. Zhang, L. Chen, M. Z. Shou, and Y. Zou, "Deep motion prior for weakly-supervised temporal action localization," *arXiv preprint arXiv:2108.05607*, 2021. 10

[158] S. Qu, G. Chen, Z. Li, L. Zhang, F. Lu, and A. Knoll, "Acm-net: Action context modeling network for weakly-supervised temporal action localization," *arXiv preprint arXiv:2104.02967*, 2021. 10

[159] C. Ju, P. Zhao, S. Chen, Y. Zhang, X. Zhang, and Q. Tian, "Adaptive mutual supervision for weakly-supervised temporal action localization," *arXiv preprint arXiv:2104.02357*, 2021. 10

[160] Y. Ge, X. Qin, D. Yang, and M. Jagersand, "Deep snippet selective network for weakly supervised temporal action localization," *Pattern Recognition*, vol. 110, p. 107686, 2021. 10

[161] R. Su, D. Xu, L. Zhou, and W. Ouyang, "Improving weakly supervised temporal action localization by exploiting multi-resolution information in temporal domain," *IEEE Transactions on Image Processing*, vol. 30, pp. 6659–6672, 2021. 10

[162] Y. Zhai, L. Wang, W. Tang, Q. Zhang, N. Zheng, and G. Hua, "Action coherence network for weakly-supervised temporal action localization," *IEEE Transactions on Multimedia*, 2021. 10

[163] A. Pardo, H. Alwassel, F. Caba, A. Thabet, and B. Ghanem, "Refineloc: Iterative refinement for weakly-supervised action localization," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 3319–3328. 10

[164] J. Li, T. Yang, W. Ji, J. Wang, and L. Cheng, "Exploring denoised cross-video contrast for weakly-supervised temporal action localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19 914–19 924. 10, 11

[165] L. Huang, L. Wang, and H. Li, "Weakly supervised temporal action localization via representative snippet knowledge propagation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3272–3281. 10, 11

[166] Y. Cheng, Y. Sun, H. Fan, T. Zhuo, J.-H. Lim, and M. Kankanhalli, "Entropy guided attention network for weakly-supervised action localization," *Pattern Recognition*, vol. 129, p. 108718, 2022. 10

[167] L. Huang, L. Wang, and H. Li, "Multi-modality self-distillation for weakly supervised temporal action localization," *IEEE Transactions on Image Processing*, vol. 31, pp. 1504–1519, 2022. 10, 11

[168] Z. Chen, H. Liu, L. Zhang, and X. Liao, "Multi-dimensional attention with similarity constraint for weakly-supervised temporal action localization," *IEEE Transactions on Multimedia*, 2022. 10

[169] J. Zhou, L. Huang, L. Wang, S. Liu, and H. Li, "Improving weakly supervised temporal action localization by bridging train-test gap in pseudo labels," in *CVPR*, 2023, pp. 23 003–23 012. 10, 11

[170] C. Ju, K. Zheng, J. Liu, P. Zhao, Y. Zhang, J. Chang, Q. Tian, and Y. Wang, "Distilling vision-language pre-training to collaborate with weakly-supervised temporal action localization," in *CVPR*, 2023, pp. 14 751–14 762. 10, 11

[171] H. Ren, W. Yang, T. Zhang, and Y. Zhang, "Proposal-based multiple instance learning for weakly-supervised temporal action localization," in *CVPR*, 2023, pp. 2394–2404. 10, 11

[172] G. Li, D. Cheng, X. Ding, N. Wang, X. Wang, and X. Gao, "Boosting weakly-supervised temporal action localization with text information," in *CVPR*, 2023, pp. 10 648–10 657. 10, 11

[173] Y. Wang, Y. Li, and H. Wang, "Two-stream networks for weakly-supervised temporal action localization with semantic-aware mechanisms," in *CVPR*, 2023, pp. 18 878–18 887. 10, 11

[174] Z. Li, Z. Wang, and Q. Liu, "Actionness inconsistency-guided contrastive learning for weakly-supervised temporal action localization," in *AAAI*, vol. 37, no. 2, 2023, pp. 1513–1521. 10, 11

[175] Y. Xu, C. Zhang, Z. Cheng, J. Xie, Y. Niu, S. Pu, and F. Wu, "Segregated temporal assembly recurrent networks for weakly supervised multiple action detection," in *AAAI*, vol. 33, no. 01, 2019, pp. 9070–9078. 10, 11, 12

[176] W. Yang, T. Zhang, X. Yu, T. Qi, Y. Zhang, and F. Wu, "Uncertainty guided collaborative training for weakly supervised temporal action detection," in *CVPR*, 2021, pp. 53–63. 10, 11

[177] X.-Y. Zhang, H. Shi, C. Li, and X. Shi, "Action shuffling for weakly supervised temporal localization," *arXiv preprint arXiv:2105.04208*, 2021. 10

[178] X.-Y. Zhang, H. Shi, C. Li, P. Li, Z. Li, and P. Ren, "Weakly-supervised action localization via embedding-modeling iterative optimization," *Pattern Recognition*, vol. 113, p. 107831, 2021. 10

[179] J. Gao, M. Chen, and C. Xu, "Fine-grained temporal contrastive learning for weakly-supervised temporal action localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19 999–20 009. 10, 11

[180] L. Huang, Y. Huang, W. Ouyang, and L. Wang, "Relational prototypical network for weakly supervised temporal action localization," in *AAAI*, vol. 34, no. 07, 2020, pp. 11 053–11 060. 10

[181] Z. Yang, J. Qin, and D. Huang, "Acgnet: Action complement graph network for weakly-supervised temporal action localization," in *AAAI*, 2022. 10

[182] Q. Liu, Z. Wang, R. Chen, and Z. Li, "Convex combination consistency between neighbors for weakly-supervised action localization," *arXiv preprint arXiv:2205.00400*, 2022. 10

[183] B. Li, B. Guo, Y. Zhu, J. Yin, and X. Ji, "Superframe-based temporal proposals for weakly supervised temporal action detection," *IEEE Transactions on Multimedia*, pp. 1–1, 2022. 10

[184] J.-T. Lee, S. Yun, and M. Jain, "Leaky gated cross-attention for weakly supervised multi-modal temporal action localization," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 3213–3222. 10

[185] Y. Zhai, L. Wang, W. Tang, Q. Zhang, N. Zheng, D. Doermann, J. Yuan, and G. Hua, "Adaptive two-stream consensus network for weakly-supervised temporal action localization," *TPAMI*, vol. 45, no. 4, pp. 4136–4151, 2022. 10, 11

[186] B. Wang, X. Zhang, and Y. Zhao, "Exploring sub-action granularity for weakly supervised temporal action localization," *TCSVT*, 2021. 9, 12

[187] T. Yu, Z. Ren, Y. Li, E. Yan, N. Xu, and J. Yuan, "Temporal structure mining for weakly supervised action detection," in *ICCV*, 2019, pp. 5522–5531. 11

[188] X.-Y. Zhang, H. Shi, C. Li, and P. Li, "Multi-instance multi-label action recognition and localization based on spatio-temporal pre-trimming for untrimmed videos," in *AAAI*, vol. 34, no. 07, 2020, pp. 12 886–12 893. 11

[189] X.-Y. Zhang, C. Li, H. Shi, X. Zhu, P. Li, and J. Dong, "Adapnet: Adaptability decomposing encoder-decoder network for weakly supervised action recognition and localization," *TNNLS*, 2020. 11, 12

[190] G. Gong, X. Wang, Y. Mu, and Q. Tian, "Learning temporal co-attention models for unsupervised video action localization," in *CVPR*, 2020, pp. 9819–9828. 11

[191] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012. 12

[192] Y.-G. Jiang, J. Liu, A. Roshan Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar. THUMOS challenge: Action recognition with a large number of classes. [Online]. Available: http://crcv.ucf.edu/THUMOS14/ 12

[193] Y. Liu, L. Wang, Y. Wang, X. Ma, and Y. Qiao, "Fineaction: A fine-grained video dataset for temporal action localization," *TIP*, vol. 31, pp. 6937–6950, 2022. 12

[194] H. Zhao, A. Torralba, L. Torresani, and Z. Yan, "Hacs: Human action clips and segments dataset for recognition and temporal localization," in *ICCV*, 2019, pp. 8668–8678. 12

[195] D. Shao, Y. Zhao, B. Dai, and D. Lin, "Finegym: A hierarchical video dataset for fine-grained action understanding," in *CVPR*, 2020, pp. 2616–2625. 12

[196] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "Youtube-8m: A large-scale video classification benchmark," *arXiv preprint arXiv:1609.08675*, 2016. 12

[197] L. Zhang, X. Chang, J. Liu, M. Luo, S. Wang, Z. Ge, and A. Hauptmann, "Zstad: Zero-shot temporal activity detection," in *CVPR*, 2020, pp. 879–888. 12

[198] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, "Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly," *TPAMI*, vol. 41, no. 9, pp. 2251–2265, 2018. 12

[199] Z. Li, L. He, and H. Xu, "Weakly-supervised temporal action detection for fine-grained videos with hierarchical atomic actions," in *ECCV*. Springer, 2022, pp. 567–584. 12

[200] M. Chen, J. Gao, and C. Xu, "Cascade evidential learning for open-world weakly-supervised temporal action localization," in *CVPR*, 2023, pp. 14 741–14 750. 12

[201] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023. 13

[202] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," *arXiv preprint arXiv:2301.12597*, 2023. 13

[203] J. Ji, K. Cao, and J. C. Niebles, "Learning temporal action proposals with fewer labels," in *ICCV*, 2019, pp. 7073–7082. 13

[204] X. Ding, N. Wang, X. Gao, J. Li, X. Wang, and T. Liu, "Kfc: An efficient framework for semi-supervised temporal action localization," *TIP*, vol. 30, pp. 6869–6878, 2021. 13

[205] X. Wang, S. Zhang, Z. Qing, Y. Shao, C. Gao, and N. Sang, "Self-supervised learning for semi-supervised temporal action proposal," in *CVPR*, 2021, pp. 1905–1914. 13

[206] B. Shi, Q. Dai, J. Hoffman, K. Saenko, T. Darrell, and H. Xu, "Temporal action detection with multi-level supervision," in *ICCV*, 2021, pp. 8022–8032. 13
[207] S. Nag, X. Zhu, Y.-Z. Song, and T. Xiang, "Semi-supervised temporal action detection with proposal-free masking," in *ECCV*. Springer, 2022, pp. 663–680. 13
[208] H. Kang, K. Kim, Y. Ko, and S. J. Kim, "Cag-qil: Context-aware actionness grouping via q imitation learning for online temporal action localization," in *ICCV*, 2021, pp. 13 729–13 738. 13
[209] Y. H. Kim, H. Kang, and S. J. Kim, "A sliding window scheme for online temporal action localization," in *ECCV*. Springer, 2022, pp. 653–669. 13
[210] F. Ma, L. Zhu, Y. Yang, S. Zha, G. Kundu, M. Feiszli, and Z. Shou, "Sf-net: Single-frame supervision for temporal action localization," in *ECCV*, 2020, pp. 420–437. 13
[211] M. Jain, A. Ghodrati, and C. G. Snoek, "Actionbytes: Learning from trimmed videos to localize actions," in *CVPR*, 2020, pp. 1171–1180. 13
[212] F. Long, T. Yao, Z. Qiu, X. Tian, J. Luo, and T. Mei, "Learning to localize actions from moments," in *ECCV*, 2020, pp. 137–154. 13
[213] J.-T. Lee, M. Jain, H. Park, and S. Yun, "Cross-attentional audio-visual fusion for weakly-supervised action localization," in *ICLR*, 2020. 13
[214] B. Wang, L. Yang, and Y. Zhao, "Polo: Learning explicit cross-modality fusion for temporal action localization," *SPL*, vol. 28, pp. 503–507, 2021. 13
[215] J. Gao, C. Sun, Z. Yang, and R. Nevatia, "Tall: Temporal activity localization via language query," in *ICCV*, 2017, pp. 5267–5275. 13
[216] J. Mun, M. Cho, and B. Han, "Local-global video-text interactions for temporal grounding," in *CVPR*, 2020, pp. 10 810–10 819. 13
[217] X. Lan, Y. Yuan, X. Wang, Z. Wang, and W. Zhu, "A survey on temporal sentence grounding in videos," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 19, no. 2, pp. 1–33, 2023. 13
[218] W. Bao, Q. Yu, and Y. Kong, "Opental: Towards open set temporal action localization," in *CVPR*, 2022, pp. 2979–2989. 13
[219] H. Alwassel, F. C. Heilbron, V. Escorcia, and B. Ghanem, "Diagnosing error in temporal action detectors," in *ECCV*, 2018, pp. 256–272. 13
[220] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *ICCV*, 2015, pp. 1422–1430. 14
[221] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *CVPR*, 2020, pp. 9729–9738. 14
[222] A. Piergiovanni, A. Angelova, and M. S. Ryoo, "Evolving losses for unsupervised video representation learning," in *CVPR*, 2020, pp. 133–142. 14
[223] G. Wang, Y. Zhou, C. Luo, W. Xie, W. Zeng, and Z. Xiong, "Unsupervised visual representation learning by tracking patches in video," in *CVPR*, 2021, pp. 2563–2572. 14
[224] R. Qian, T. Meng, B. Gong, M.-H. Yang, H. Wang, S. Belongie, and Y. Cui, "Spatiotemporal contrastive video representation learning," in *CVPR*, 2021, pp. 6964–6974. 14
[225] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, 2020. 14
[226] D. Bogdanov, S. Laur, and J. Willemson, "Sharemind: A framework for fast privacy-preserving computations," in *European Symposium on Research in Computer Security*, 2008, pp. 192–206. 14
[227] D. Demmler, T. Schneider, and M. Zohner, "Aby-a framework for efficient mixed-protocol secure two-party computation." in *NDSS*, 2015. 14
[228] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*, 2017, pp. 1273–1282. 14
[229] C. E. Shannon, "A mathematical theory of communication," *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948. 14

Binglu Wang (M'21) received the Ph.D. degree in Control Science and Engineering with the School of Automation at Northwestern Polytechnic University , Xi'an, China, in 2021. He is currently a Post-doctoral with the Beijing Institute of Technology, Beijing, China. His research interests include Computer Vision, Digital Signal Processing and Deep Learning.

Yongqiang Zhao (M'05) received the B.S., M.S., and Ph.D. degrees in control science and engineering from the Northwestern Polytechnic University,Xi'an, China.From 2007 to 2009, he was as a Post-Doctora Researcher with McMaster University, Hamilton, ON, Canada, and Temple University, Philadelphia,PA, USA. He is currently a Professor with the Northwestern Polytechnical University. His research interests include polarization vision, hyperspectral imaging, and pattern recognition.

Le Yang received his B.E. and Ph.D. degree from Northwestern Polytechnical University in 2016 and 2022. His research interests include video analysis and weakly supervised learning.

Teng Long was born in Fujian, China, in 1968. He received the M.S. and. Ph.D. degrees in electrical engineering from the Beijing Institute of Technology, Beijing, China, in 1991 and 1995, respectively. He was a Visiting Scholar with Stanford University, California, in 1999, and University College London, in 2002. He has been a Full Professor with the Department of Electrical Engineering, Beijing Institute of Technology, since 2000. He has authored or co-authored more than 300 articles. His research interests include synthetic aperture radar systems and real-time digital signal processing, with applications to radar and communication systems. Dr. Long is a Fellow of the Institute of Electronic and Technology and the Chinese Institute of Electronics. He was the recipient of many awards for his contributions to research and invention in China. He has been a member of the Chinese Engineering Academy since 2021.

Xuelong Li (M'02-SM'07-F'12) is a full professor with School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, P.R. China.