Artificial Intelligence Review (2024) 57:26
https://doi.org/10.1007/s10462-023-10650-w

Overview of temporal action detection based on deep learning

Kai Hu[1,2] · Chaowen Shen[1,2] · Tianyan Wang[1,2] · Keer Xu[1,2] · Qingfeng Xia[3,4] · Min Xia[1,2] · Chengxue Cai[1,2]

Accepted: 19 December 2023 / Published online: 1 February 2024

Abstract
Temporal Action Detection (TAD) aims to accurately capture each action interval in an untrimmed video and to understand human actions. This paper comprehensively surveys the state-of-the-art techniques and models used for TAD task. Firstly, it conducts comprehensive research on this field through Citespace and comprehensively introduce relevant dataset. Secondly, it summarizes three types of methods, i.e., anchor-based, boundary-based, and query-based, from the design method level. Thirdly, it summarizes three types of supervised learning methods from the level of learning methods, i.e., fully supervised, weakly supervised, and unsupervised. Finally, this paper explores the current problems, and proposes prospects in TAD task.

Keywords Temporal action detection · Video localization · Proposal generation · CiteSpace

Abbreviations

AFSD	Anchor-free saliencybased detector
AFSD	Anchor-free saliencybased detector
AP	Average precision
AR	Average recall
BGCN	Boundary graph convolutional network
BM	Boundary-matching
BSN	Boundary sensitive network
BSP	Boundary-sensitive pretex
C3D	Convolutional 3D
CAS	Class activation sequence
CFAD	Coarse-to-fine action detector
CLAP	Contrastive language-action pre-training
CNN	Convolutional neural network
CPD	Coarse pyramidal detection
CPN	Contextual proposal network

Chaowen Shen, Tianyan Wang, Keer Xu, Qingfeng Xia, Min Xia and Chengxue Cai have contributed equally to this work.

Extended author information available on the last page of the article

CW	Class-wise foreground classification branch
DCNN	Dilated convolution neural network
DT	Dense trajectories
FGD	Fine-grained detection
GTAN	Gaussian temporal awareness networks
GTE	Global temporal encoder
HOG	Histogram of oriented gradient
I3D	Inflated 3D ConvNet
IMU	Inertial measurement units
LGTE	Local–global temporal encoder
LRCN	Long-term recurrent convolutional network
LSTA	Long short-term attention
LSTM	Long short-term memory
LTE	Local temporal encoder
MAP	Mean average precision
MIL	Multi-instance learning
NCE	Noise contrastive estimation
NMS	Non maximum suppression
OIC	Outer-Inner-Contrastive
P3D	Pseudo-3D
PAL	Pseudo action localization
PFGCN	Proposal features graph convolutional network
RCL	Recurrent continuous localization
R-CNN	Region-based convolutional neural networks
RNN	Recurrent neural network
RPD	Refined pyramidal detection
RPN	RegionProposal network
SFTP	SuperFrame-based temporal proposal
SSAD	Single shot action detector
SSD	Single shot multiBox detector
STIP	Space-time interest points
TAD	Temporal action detection
TAL	Temporal action localization
TBR	Temporal boundary regressor
TSCN	Two-stream consensus network
TSN	Temporal segment network
UGPT	Uncertainty-guided probabilistic transformer
ViT	Vision transformer

1 Introduction

In recent years, with the development of multimedia technology and the rapid popularization of digital equipment (Graziani et al. 2022), the amount of Internet video data has grown significantly. Therefore, how to deal with these multimedia data efficiently and accurately has become a hot research topic (Le et al. 2021). The purpose of video understanding is to automatically identify and parse the content of a video using intelligent analysis technology. Given the success of deep learning in image processing and detection,

researchers have introduced deep learning methods to the field of video understanding (Hutchinson and Gadepally 2021).

Computer vision tasks relating to human action mainly include action recognition (Tran et al. 2015; Hu et al. 2022b; Wang et al. 2022b), action prediction (Kong and Fu 2022), and temporal action detection (Xia and Zhan 2020). Important achievements have been made in the field of action recognition in respect of facial recognition (Li et al. 2022b) and video surveillance. Researchers typically use edited videos for action recognition with only one action. Therefore, action recognition only needs to classify the action without detecting the duration of the action. However, most videos in real life are untrimmed and may contain multiple instances of actions in different categories, each with an unknowable time boundary and duration. In 2017, during the ActivityNet Big Action Recognition Challenge organized by CVPR (Ghanem et al. 2017), video understanding was divided into five branch tasks: untrimmed video classification, trimmed action recognition, temporal action proposals, temporal action localization (temporal action detection), and dense-captioning events in videos. Temporal action detection refers to locating action instances and identifying action categories in untrimmed videos, which are more complex tasks than action recognition. Therefore, in this paper we focus on temporal action detection. Figure 1 shows an example of long-jump temporal action detection, where the start and end times of the movement are obtained by localization. The target of detection is a predefined action category, and the time interval of other activities that do not belong to this group of actions is called the time background.

Standardization efforts in temporal action detection date back to 2007, when Ke et al. (2007) used handcrafted feature methods to detect specific actions in fixed-camera kitchen cooking videos. With the emergence of the THUMOS-14 dataset, the temporal action detection task has been further developed. In 2014, Oneata et al. (2014a) and Wang et al. (2014) used DT features and single-frame CNN features respectively to generate candidate segments of specific sizes through sliding windows, and built a framework for temporal action detection.Later, Yuan et al. (2016) proposed a temporal action detection algorithm using iDT features, which uses iDT features to extract pyramid score distribution features (PSDF) to describe actions in videos.

However, traditional feature extraction methods have time and storage overhead in sequential action detection. In order to solve this problem, Shou et al. (2016) introduced the anchor mechanism in 2016 and proposed a multi-stage SCNN, which combines the sliding window and the proposal generation network to effectively perform action detection. However, it still requires a lot of computation when processing action instances of

Fig. 1 Example of temporal action detection's execution of the long jump

different durations. Therefore, researchers have also proposed some improved methods based on the anchor mechanism, such as boundary-based temporal action detection algorithms TAG (Xiong et al. 2017) and SSN (Lin et al. 2018), and methods that use action probability distribution curves to improve the confidence of candidate segments, which can provide flexible temporal boundaries make up for the shortcomings of anchor-based methods in terms of precise action boundaries, but may generate proposals with relatively low confidence.Moreover, query-based temporal action detection methods have garnered significant interest. These approaches model action instances as a collection of learnable action queries, eliminating the constraints of manually designing anchor points and boundaries, and offering the benefit of simplifying the computational pipeline.

All in all, multiple algorithms have made significant progress in the domain of temporal action detection. From the earliest hand-engineered feature techniques to deep learning-based methods, researchers have persistently introduced innovative algorithms and models to tackle numerous challenges in temporal action detection. These methods not only improve the accuracy and robustness of the technology, but also bring a lot of value to practical applications. Especially in the areas of anomaly detection, teaching video analysis, and sports video analysis, sequential action detection has achieved good practical results. For example, it can automatically identify and locate abnormal events in videos, which greatly guarantees safety monitoring; in teaching video analysis, it can automatically segment and locate key actions in videos, which greatly facilitates learners' learning process. In the future, this field will continue to flourish, further expanding its influence across a wide range of application scenarios.

However, in real-world situations, temporal action detection encounters a multitude of obstacles and unresolved matters, the key hurdles being as follows:

(1) Time information. Because of the one-dimensional time series information, static image information cannot be used for temporal action detection. It must be combined with time series information;
(2) The boundary is unclear. Unlike in object detection, the boundary of the target is usually very clear, and a clearer boundary box can be drawn for the target. However, there may be no reasonable definition of the exact timeframe for the operation. Therefore, it is impossible to provide exact boundaries at the beginning and end of an action;
(3) The time span is large. The span of time action segments can be very large. For example, waving a hand takes only a few seconds, while rock climbing or cycling can last several minutes. The task spans differ in length, making it extremely difficult to extract schemes from them. In addition, in an open environment, there are problems of multiscale, multi-target, camera action, etc.

In the previous review by Xia and Zhan (2020) in 2020, time series action detection was divided into single-stage and two-stage types. The one-stage method was briefly described as the generation of candidate proposals and the classification of movements simultaneously. The two-stage approach was described as processing the candidate proposals first, then classifying and regressing the actions. This classification method is simple, ignores the internal design ideas and characteristics of the model, and only works around the process to perform a simple induction, which makes it suitable for beginners to learn. In addition, most of the learning materials focus on fully supervised learning. Although the current performance of weakly supervised learning is weak, weakly supervised learning will be closer to reality, and its development over

the past two years has been relatively rapid; it is a method that cannot be ignored and should be introduced in detail. In 2022, Baraka and Mohd Noor (2022) published a review on weak supervision, which introduced concepts, strategies, and technologies related to weak supervision in detail. Weak supervision was divided into two methods, bottom-up and top-down, which were not comprehensively classified or detailed in the introduction to their paper.

In this review we use a new classification method, namely multi-instance learning and direct localization, to introduce weak supervision and the development of unsupervised learning. Full and limited supervision methods are equally important, and we consider them in this paper. The contributions of this review are as follows:

(1) The literature in recent years is summarized and updated comprehensively in respect of each stage of temporal action detection;
(2) Video feature extraction methods are summarized in detail using three learning methods;
(3) The system model is horizontally divided into three categories (anchor-based, boundary-based, and query-based) according to the implementation method, and vertically divided into full-supervision and limited-supervision learning methods. This review may help scholars to understand the task of temporal action detection comprehensively.

The structure of the review is as follows. Section 2 introduces the relevant background, which can assist beginners in understanding the basic concept of temporal action detection tasks, including the form of the dataset and the explanation of common nouns. Section 3 also presents a CiteSpace analysis. With the help of CiteSpace software, the research hotspots and research areas of this task are presented visually, and the topic of this paper is objectively and comprehensively explained using keyword co-occurrence, topic clustering, and other methods. Section 4 introduces research in respect of video feature extraction, which is divided into traditional methods and deep learning methods. Video feature extraction must be introduced as a public step of video task processing. In this review we introduce feature extraction by means of CNN, RNN, and transformers. Section 5 introduces algorithms based on algorithm structure; these can be divided into three types according to the design methods: anchor-based, boundary-based and query-based. Section 6 introduces algorithms based on the supervision mode, focusing on weak supervision, and provides a general introduction to full supervision and no supervision. To enable readers to obtain a deeper and more comprehensive understanding of temporal action detection, Sects. 5 and 6 introduce temporal action detection in horizontal and vertical ways, from the design method to the learning method, using two different research routes to make the structure of the paper more rigorous. Sections 7 and 8 present conclusions and prospects. Figure 2 shows the structure diagram of the article.

2 Background

This section introduces some relevant information in respect of temporal action detection. Section 2.1 introduces the basic concepts and commonly used evaluation methods, and Sect. 2.2 reviews the video datasets widely used in temporal action detection tasks.

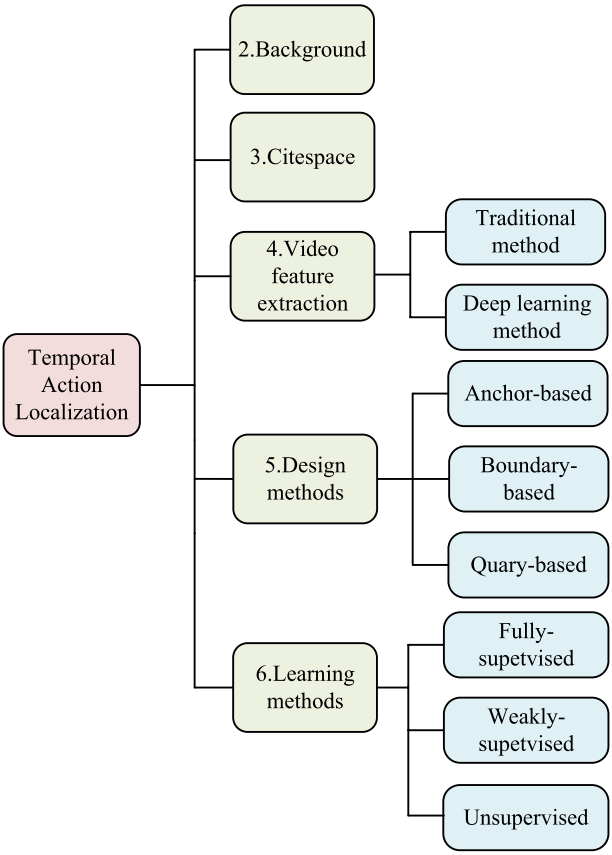


Fig. 2 Overall block diagram

2.1 Basic concepts and evaluation metrics

Definition 1 Temporal action detection: Temporal action detection can be regarded as image target detection with a time sequence channel, aiming at dividing and identifying the action intervals in untrimmed video, then outputting the start and end time of each action and the action category. It can be expressed as:

$$W_x = \left\{\Psi_a = (t_s, t_e, l_a)\right\}_{a=1}^{N} \tag{1}$$

where ω_a is a group of action examples; N indicates the number of actions in this group of action instances; Ψ_a the A-th action instance; t_s,t_e,l_a indicate the start time, end time, and corresponding label of the action instance, respectively; labels l_a belong to $\{1, 2, 3, ..., C\}$; and C is the category in the dataset. One can select different annotation information in the action instance when dealing with different learning methods, as shown in Fig. 3.

Definition 2 Temporal proposals: Temporal proposal P is a segment that may contain or partially contain an action; the comment information for each P includes t_s,t_e,l_a and the

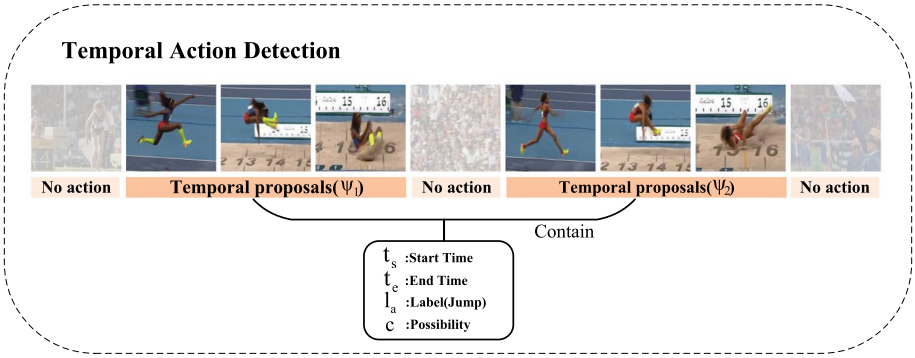


Fig. 3 Examples of temporal action detection and temporal proposals, proposals(instances of actions) containing information such as start time, end time, label category, and confidence

confidence score c; the confidence score is the probability of predicting that P contains an instance of the action. Therefore, P can be represented as (t_s, t_e, l_a, c), as shown in Fig. 3.

Definition 3 Action classification: After the generation of the temporal proposal, the proposal needs to be fed into the action classifier for action classification. Most temporal action detection models use existing action classifiers for classification.

Definition 4 Video feature extraction: For untrimmed video, it is difficult to input the whole video into the encoder for feature extraction; therefore, the video needs to be segmented and input into the pre-trained video encoder for feature extraction. Each video can be represented by a series of visual features that are further processed for action detection. Common visual encoders are two-stream (Simonyan and Zisserman 2014), I3D (Carreira and Zisserman 2017), C3D (Tran et al. 2015), TSN (Wang et al. 2016b), R(2 + 1)D (Tran et al. 2018), and P3D (Qiu et al. 2017). These are explained in Sect. 4.

Definition 5 Evaluation metrics for temporal action proposal: The commonly used evaluation criterion for this task is average recall (AR). Intersection over union (IOU) thresholds are also set. For the two most widely used public datasets, ActivityNet−1.3 and THUMOS14, the thresholds of IOU are generally set as [0.5,0.05,0.95] and [0.5,0.05,1]. To more accurately evaluate the relationship between the recall rate and the average number of proposals, the relationship curve AR@AN between the average recall rate (AR) and the average number of proposals (AN) is generally adopted.

Definition 6 Evaluation metrics for temporal action detection(TAD): For TAD tasks, mean average precision (MAP) is used as the average standard, and average accuracy (AP) is calculated in respect of the class of pairs of actions. For the two most widely used public datasets, ActivityNet−1.3 and THUMOS14, the thresholds of IOU are generally set as [0.5,0.75,0.95] and [0.3,0.4,0.5,0.6,0.7]. The higher the IOU threshold, the more difficult an object is to detect. IOU with different threshold values is usually selected in experiments to comprehensively test the model's performance.

Definition 7 Dataset Preprocessing: Dataset preprocessing consists of two main steps: data collection and data segmentation. In order to collect data, researchers can use publicly available datasets, like UCF101(Soomro et al. 2012) and AVA (Gu et al. 2018) mentioned in Sect. 2.2, or they can obtain data through specialized video and sensor collection equipment. Moreover, data segmentation is a vital aspect of temporal action detection dataset preprocessing. This process involves splitting lengthy videos or continuous time series data into shorter segments or time windows so that action instances can be processed individually. Popular segmentation methods encompass: fixed-time interval segmentation, motion boundary detection-based segmentation, and motion event-based segmentation. Figure 4 shows the flow chart of dataset preprocessing, drawn with Definition8, Definition9 and Definition10.

Definition 8 Labeling and Annotation of Datasets: Accurate labeling and annotation are essential for providing vital information needed for model training and evaluation. Dataset labeling and annotation involve time annotation, category annotation, and bounding box annotation:

(1) Time annotation is utilized to mark the start and end timestamps of an action, enabling the chronological identification of events within a video or time series.
(2) Category annotation entails assigning a specific category or group of categories to each action instance, indicating the action's type or class.
(3) Bounding box annotation indicates the spatial location of an action within the video or image. Bounding boxes can be rectangles or polygons that encompass the area where the action takes place.

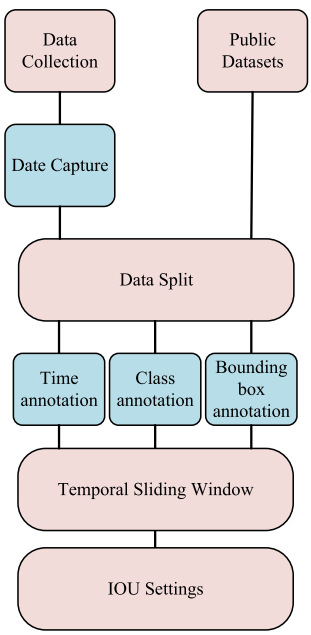


Fig. 4 Flow chart of dataset preprocessing

For efficient and effective annotation, professional annotation tools are available. These tools offer visual interfaces and interactive features, allowing researchers to effortlessly annotate time, category, and bounding boxes. Widely used labeling tools include Labelbox, OpenLabeling, and others.

Definition 9 Temporal sliding window: The temporal sliding window technique is often used in action segmentation during dataset preprocessing for temporal action detection. Its purpose is to detect actions within video sequences. The fundamental idea is to slide a fixed-size window along the timeline, carrying out feature extraction and action detection on the data within the window at each position. By moving the window across the timeline, continuous action detection can be performed on actions at different locations in the video sequence,Regarding the settings of the temporal sliding window, the following description can be given:

$$\begin{aligned}(X, window_{size}, stride) &= X[i, i + window_{size}] \mid i \\ &= 0, stride, 2stride, \ldots, len(X) - window_{size}\end{aligned} \tag{2}$$

Among them, X represents the original video sequence, $window_{size}$ represents the window size, and Stride represents the step size of the sliding window. On the time series X, starting from an index position i, slide the window with a step size of Stride to obtain a series of fixed-size subsequences. Figure 5 shows a schematic of the temporal sliding window, where $window_{size} = 3$, $Stride = 3$.The window slides along the video frames at a rate of overlapping 1 frame each time.

Definition 10 Overlap rate setting:In temporal action detection, the sliding window setup is usually configured according to Definition 9. In order to capture as much information from the video as possible, the sliding window moves across the entire video, thereby generating a certain overlap rate. This refers to the ratio of frames shared between two consecutive windows.

Typically, the stride (or moving speed) of a sliding window determines the overlap rate between windows. Establishing an optimal stride size is crucial, as it assists researchers in capturing vital actions without compromising on the efficacy of action detection. Generally speaking, if the stride size is set small, the overlap rate will be large, and there will be a large number of shared video frames between two adjacent windows. As a result, although the accuracy of action detection is increased, it also increases processing requirements and computational costs. If the stride size is set larger, the overlap rate will be lower and fewer video frames will be shared between two adjacent windows. However, the accuracy

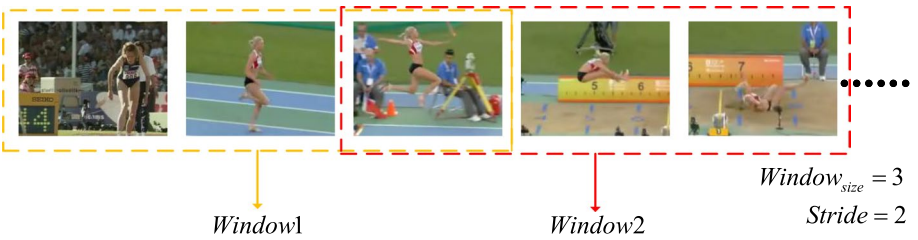


Fig. 5 Flow chart of dataset preprocessing

of action detection may be reduced because some important short-term actions may be missed.

In temporal action detection, the overlap rate of the sliding window is generally set at 50%-75%. The specific settings depend on various factors, including the availability of computing resources, the specific characteristics of the actions that need to be detected, and the characteristics of the dataset. This requires a lot of experimentation and optimization work.

2.2 Datasets

There are many common datasets in the field of video understanding (Pareek and Thakkar 2021). Comparison of the performance of different models requires datasets as the carrier. There are two problems with most datasets in use today: first, compared with the rich types of actions in human life, the number of categories in most datasets is very small, such as in KTH (Schuldt et al. 2004), UCF sports (Rodriguez et al. 2008), Weizmann (Weinland et al. 2007), etc.; second, the source of many datasets is not real enough and lacks the unique interference that occurs in the real environment. For example, HOHA (Marszalek et al. 2009) and UCF Sports are composed of professionally photographed teams or non-real scenes taken from movie clips. However, as action-capture systems mature and crowd-sourced tagging services improve, these problems will become easier to mitigate. Next, we introduce some mainstream temporal action detection datasets, as listed in Table 1.

(1) UCF101 (Soomro et al. 2012)

UCF101, an action recognition dataset of realistic action videos, contains 101 action categories and a total of 13,320 videos, with a total duration of 27 h. As an extension of the UCF50 dataset, UCF101 has enriched the categories of movements, including five categories: human-object interaction, simple body movements, human-person interaction, playing musical instruments, and sports. As the first large-scale action recognition dataset, UCF101 has a representative position in video datasets. THUMOS '14, MEXaction2, and other datasets all refer to some of the videos in UCF101. In addition, UCF Sports, UCF11 (Liu et al. 2009), UCF50 (Reddy and Shah 2013), and UCF101 are four action datasets produced by UCF in chronological order. In this order, each dataset contains data from the previous dataset (Fig. 6).

(2) HMDB51 (Kuehne et al. 2011)

Previously, researchers had been working on databases of still images collected on the Internet, but the action-recognition datasets were far below average. Like the previously popular KTH, Weizmann, and IXMAS (Weinland et al. 2007) datasets, a common feature of these datasets is that there are only a small number of occlusion objects and a limited number of complex actors in the video clip environment, meaning they do not adequately represent the complexity and richness of the real world. Moreover, the recognition rate of such datasets is often very high. Therefore, to promote the sustainable development of action recognition and improve the richness and complexity of datasets, Kuehne et al. proposed HMDB51. The HMDB51 dataset is a small and easy-to-use human movement dataset, containing 51 action categories and a total of 6849 clips including manual annotation. There are five action categories: general facial action, facial action and object action,

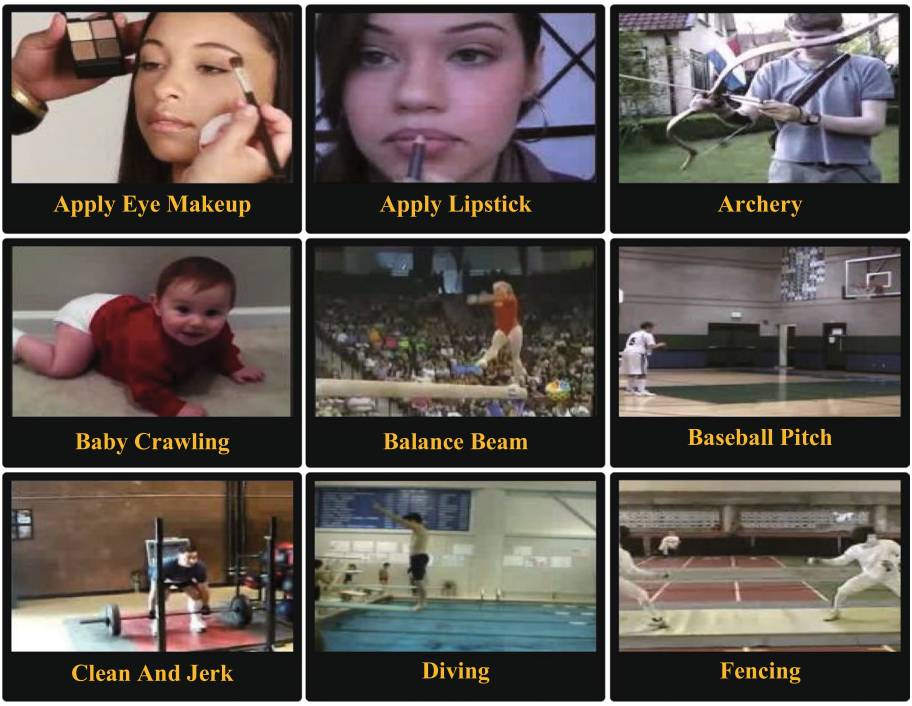


Fig. 6 UCF101 dataset

general body action, human interaction, and human action. Videos were obtained from a variety of sources, from digitized movies to YouTube (Fig. 7).

(3) THUMOS'14 (Jiang et al. 2014)

THUMOS'14 originated from the THUMOS Challenge 2014. This dataset contains many untrimmed videos of human behaviour in real-world environments, and the ability to predict activity in untrimmed video sequences can be evaluated. The main tasks are action recognition and sequential action detection. Currently, most papers use this dataset for testing and evaluation. In the action recognition task, the training set contains 13,320 trimmed videos, covering 101 action categories. The validation set contains 1010 untrimmed videos, while the test set contains 1574 untrimmed videos. For the temporal action detection task, the training set contains 213 trimmed videos, which contain 20 action categories. The validation set contains 200 untrimmed videos, while the test set also contains 1574 untrimmed videos. In addition, THUMOS-14 has been further developed in 2015's THUMOS-15, containing more than 430 h of video data, which is about 70% larger than THUMOS-14. This makes it a more challenging video action dataset.

(4) ActivityNet (Caba Heilbron et al. 2015)

While there has been an explosion in video data, with more than 300 h of video uploaded to YouTube every minute, there have been no corresponding advances in recognizing and

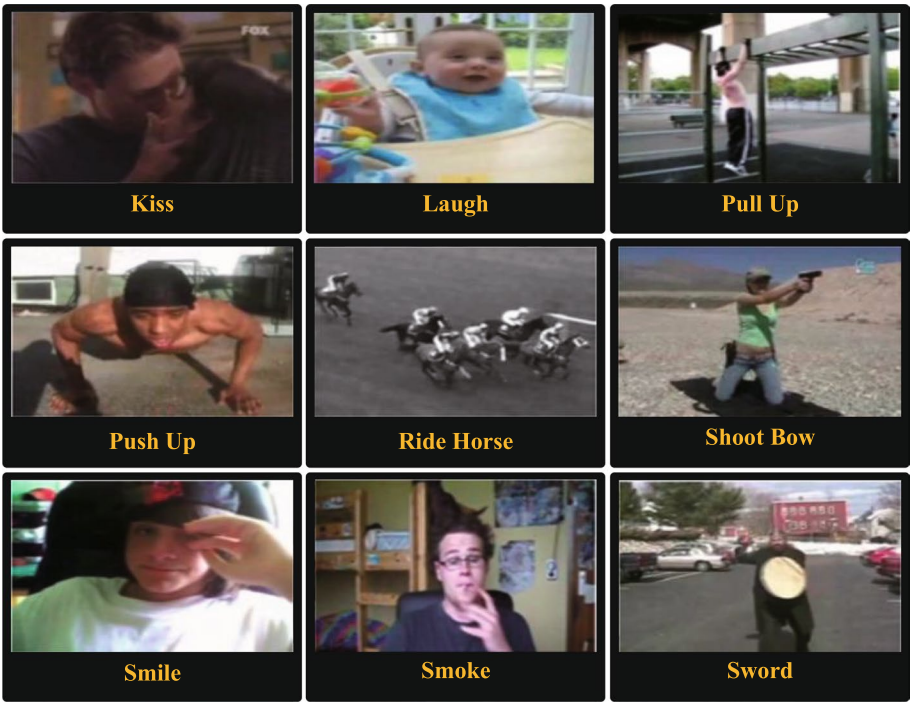


Fig. 7 HMDB51 dataset

understanding human action tasks. Most datasets ignore the vast variability in execution styles; the complexity of visual stimuli in terms of camera movement, background clutter, and changes in viewpoint; and the level of detail and amount of activity that can be identified. For example, UCF Sports (Rodriguez et al. 2008) and Olympic Sports (Niebles et al. 2010) increase the complexity of movements by focusing on highly defined physical activities. Another significant limitation is that most computer vision algorithms for understanding human activity are based on datasets covering a limited number of activity types. In fact, existing databases tend to be specific and focus on certain types of activities. To address the limitations of this dataset, Heilbron et al. proposed a flexible framework for capturing, annotating, and segmenting online video, also known as ActivityNet. The ActivityNet dataset is also used for a large-scale behavior recognition contest; the dataset consists of 27,801 videos, including 13,900 training videos, 6950 validation videos, and 6950 test videos. ActivityNet offers 203 activity categories in the current release, with an average of 137 untrimmed videos per category and 1.41 activity instances per video for 849 h.

(5) AVA (Gu et al. 2018)

The AVA dataset is a spatiotemporal detection dataset containing 430 15-minute videos tagged with 80 action categories, 1.58 million action labels, and 81,000 action tracks. To increase the diversity of the dataset, the producers cut 15-minute clips into 897 overlapping 3-second movie clips in one-second steps. The 430 videos were divided into 235 training videos, 64 verification videos, and 131 test videos. The AVA dataset is a new annotated

Table 1 Introduction to common databases

Dataset	Year	Behavior category	Number of videos	Dataset duration(hours)	Resources
UCF101 Soomro et al. (2012)	2012	101	13320	27	https://www.crcv.ucf.edu/research/data-sets/ucf101/
HMDB51 Kuehne et al. (2011)	2011	51	6849	–	https://www.serrelab.clps.brown.edu/resource/hmdb-a-large-human-motion-database/dataset
THUMOS'	2014	101	18394	–	https://www.crcv.ucf.edu/THUMOS14/download.html
ActivityNet (Caba Heilbron et al. 2015)	2015	203	27801	849	http://activity-net.org/
Charades Sigurdsson et al. (2016)	2016	157	27847	–	https://prior.allenai.org/projects/charades
AVA Gu et al. (2018)	2018	80	430	108	https://research.google.com/ava/
MEVA Corona et al. (2021)	2020	37	–	9300	https://mevadata.org/
MoVi Ghorbani et al. (2020)	2020	20	–	32.6	https://www.biomotionlab.ca/movi/
FineAction Liu et al. (2022b)	2021	106	16732	–	https://deeperaction.github.io/datasets/fineaction.html

video dataset that is human-centric, sampled at a frequency of 1Hz, and framed for each person. The side label of the boundary box is the actor's action, and the interaction between the objects is generated. In the AVA dataset, the actions of all persons are marked in the keyframe, but the result is an uneven class of actions of the Zipf law type. The action recognition model should be based on the real long-tail action distribution (Horn and Perona 2017) rather than on an artificially balanced dataset.

(6) MEVA (Corona et al. 2021)

Datasets used for action recognition often fail to meet public safety community requirements, for example AVA, moments in time (Monfort et al. 2019), and YouTube-8 M (Abu-El-Haija et al. 2016). These datasets present short, high-resolution video specificities centered on the activity of interest in both time and space. In real life, more datasets are required with an actual spatial scope. Multiview extended video with activities (MEVA) is a new dataset for human action recognition. MEVA comprises more than 9300 h of continuous uncut video that contains spontaneous background activity. In this dataset, there are 37 kinds of actions, spanning 144 h in total, and actors and props are framed with borders. In addition, about 100 actors were gathered to perform scripted scenes and spontaneous background activities in a gated and controlled venue over three weeks, and video was collected in various ways so that indoor and outdoor views overlapped or did not overlap.

(7) MoVi (Ghorbani et al. 2020)

Ghorbani et al. introduced a new human action and video dataset, MoVi, which will soon be publicly available. It comprises data from 60 female and 30 male actors performing 20 predefined daily and motor movements and one optional move. During the five-round capture process, the same actors and actions were recorded using different hardware systems, including optical action capture systems, cameras, and inertial measurement units (IMUs). In some capture rounds, the actors were recorded in their natural clothing, while in other rounds, they wore very little. The dataset contains 9 h of action capture data, 17 h of video data from four different angles, including a handheld camera, and 6.6 h of IMU data. Ghorbani et al. describe how the dataset was collected and post-processed and discuss examples of potential research that could be achieved with the dataset.

(8) Charades (Sigurdsson et al. 2016)

Computer vision technology can help people in their daily lives by finding lost keys, watering plants, or reminding people to take their medicine. To accomplish these tasks, researchers need to train computer vision methods from real and diverse examples of everyday dynamic scenarios. Sigurdsson et al. proposed a novel Hollywood family approach to collecting such data. Instead of shooting videos in a lab, Sigurdsson et al. ensured diversity by distributing and crowdsourcing the entire video creation process, from scripting to video recording and annotation. Following this process, the authors collated a new dataset, i.e., Charades. Hundreds of people recorded videos and went about their daily leisure activities at home. The dataset consists of 9848 annotated videos, with an average length of 30 s, showing the activities of 267 people across three continents, with more than 15 percent featuring more than one person. Multiple free text descriptions, action labels, action intervals, and interaction object classes annotate each video in the dataset. Users

can employ this wealth of data to evaluate and provide baseline results for multiple tasks, including action recognition and automatic description generation. The dataset's authenticity, diversity, and randomness will present unique challenges and new opportunities to the computer vision community.

(9) FineAction (Liu et al. 2022b)

Temporal action detection is an important and challenging problem in video comprehension. However, most existing TAD benchmarks are based on the coarse-grained nature of the action class. This presents two major limitations for this task. First, the rough action level causes the location model to over-adapt to high-level contextual information while ignoring the atomic action details in the video. Secondly, the rough action class usually leads to the fuzzy annotation of the time boundary, which is unsuitable for temporal action detection. To address these issues, Liu et al. developed a new large-scale fine-grained video dataset called FineAction for temporal action detection. FineAction contains 103K time instances of 106 action categories, annotated in 17K untrimmed videos. Due to the rich diversity of the fine motor class, the intensive annotation of multiple instances, and the concurrent actions of different classes, the fine motor class provides new opportunities and challenges for temporal action detection. In order to benchmark FineAction, Liu et al. systematically examined the performance of several popular TAD methods and analyzed in depth the impact of short-time and fine-grained instances of TAD.

3 Citespace analysis

Visualization of scientific knowledge based on social networks and graph theory comprises a new field of bibliometric methods. CiteSpace (Chen 2004, 2006, 2013; Chen et al. 2010) has received extensive attention worldwide due to its advanced and powerful functions. Therefore, CiteSpace was used in this study to analyze TAD tasks visually. We used CiteSpace (5.3.R11) to visualize the data. We established parameters, including time slices (annual slices were used for co-author analysis and keyword co-occurrence), keyword sources (title, abstract, author keywords, and keyword plus), and node types (author, institution, country, cited reference, cited author). Literature analysis based on CiteSpace can identify the research content and hotspots in a certain field more conveniently and quickly.

Since deep learning technology was introduced into the field of video understanding, the task of TAD has developed rapidly. In this study, 1326 articles relating to TAD from the past 12 years were retrieved via a Web of Science search for temporal action detection (TAD) and temporal action localization (TAL). As shown in Fig. 8, a keyword heatmap was created. The larger the circle, the more times the keyword appears. The round layer from the inside out represents the past to the present, and the redder the layer, the more attractive popular it is. The figure shows that keywords with more frequency have larger circles, while keywords with less frequency have smaller circles and are not displayed. From the cluster graph, it can be seen that "action recognition", "feature extraction", and "temporal action localization" are prominent. This indicates that TAD is always active in these research fields. In the figure, "location awareness" and "proposal network" are darker keywords, indicating that they appeared later; researchers need to pay more attention to these. In addition, with the development of TAD, tasks such as action prediction and sound localization have been further developed.

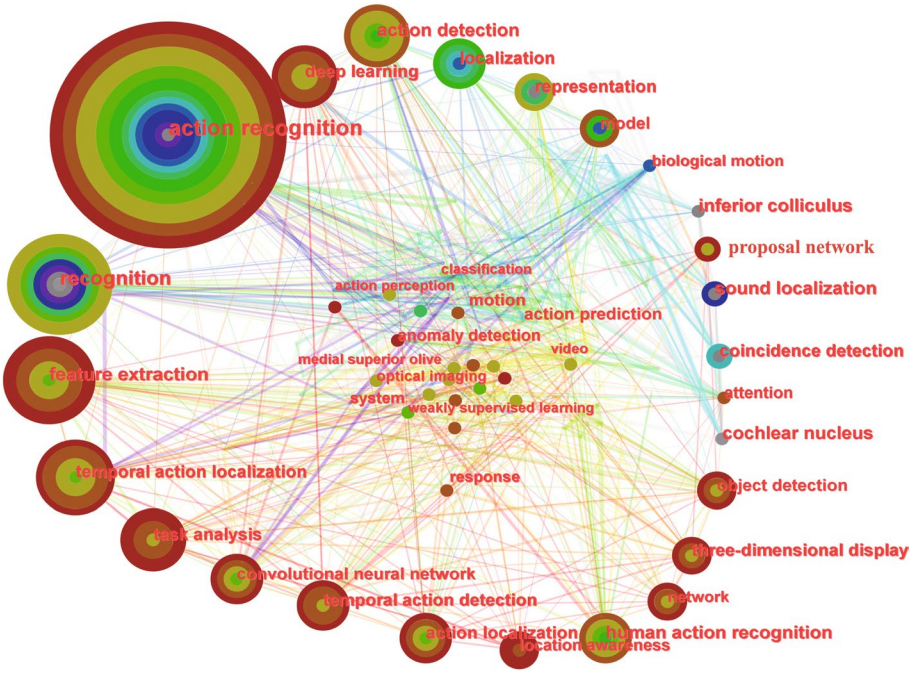


Fig. 8 Hot words in the field of temporal action detection

Figure 9 shows the time sequence diagram of the occurrence of high-frequency keywords. The first 11 keywords with high frequency from 2014 to 2022 were counted. "Strength" represents the strength of keywords. The higher the value, the more times

Top 17 Keywords with the Strongest Citation Bursts

Keywords	Year	Strength	Begin	End	2014 - 2022
localization	2014	4.1	**2014**	2018	
sound localization	2014	3.66	**2014**	2016	
motion	2014	3.12	**2014**	2019	
brain stem	2014	2.83	**2014**	2016	
cochlear nucleus	2014	1.82	**2014**	2016	
model	2014	3.44	**2015**	2018	
integration	2014	2.58	**2015**	2018	
inferior colliculus	2014	2.29	**2015**	2016	
action representation	2014	1.54	**2015**	2017	
representation	2014	2.81	**2016**	2018	
histogram	2014	3.32	**2017**	2018	
pose	2014	2.05	**2018**	2019	
classification	2014	1.87	**2018**	2019	
human action recognition	2014	1.51	**2018**	2020	
recognition	2014	1.5	**2019**	2020	
three-dimensional display	2014	2.4	**2020**	2022	
temporal action proposal generation	2014	1.86	**2020**	2022	

Fig. 9 Time sequence diagram of high-frequency keywords

the keyword has been cited. The line on the right is the timeline from 2014 to 2022. “Begin” indicates the time when the keyword first appears. “Begin” and “End” show the activity of the keyword during the year. It can be seen from the figure that words such as "localization" and "recognition" are hot topics for temporal action detection. Keywords such as "human action recognition" and "temporal action proposal generation" are also the focus of this paper. The keywords in Fig. 8 complement each other and make up for the ambiguity of the time information in Fig. 8. Through these keywords, readers can better understand the hotspot direction in respect of TAD.

With the development of deep learning, increasing numbers of scholars and teams have begun to study TAD, and they jointly promote the progress and development of TAD algorithms. In this study we analyzed 1326 articles from 2010 to 2022 (data obtained from Web of Science); the author contribution graph, as shown in Fig. 10, was obtained, in which the larger the font, the more papers have been published by the author, and the greater their influence. WANG L designed the first video-level framework to learn video representation and created the most classic TSN (Wang et al. 2016b) model. WANG L was also the first to apply action detection to weak supervision and designed UntrimmedNet (Wang et al. 2017), making remarkable contributions to the field. Zhao Y introduced the idea of image detection into time sequence action detection for the first time and designed an SSN (Zhao et al. 2017) network with an obvious improvement effect. Liu J mainly studied action recognition in TAD, involving the field of action prediction.

We retrieved 1418 articles in respect of TAD and their citation frequency through a Web of Science search for TAD and TAL. In Fig. 11, the blue bar chart shows the number of posts, and the red line shows the frequency of citations. As shown in Fig. 11, the number of papers published in the field of TAD is rising, and the frequency of citations is increasing rapidly. The drop in the 2022 figures is because the 2022 figures are only available through August. In 2012, the number of papers was relatively small, the field had not been fully developed, and there were fewer researchers. After 2013, there was

Fig. 10 Main contributors of papers in the field of temporal action detection

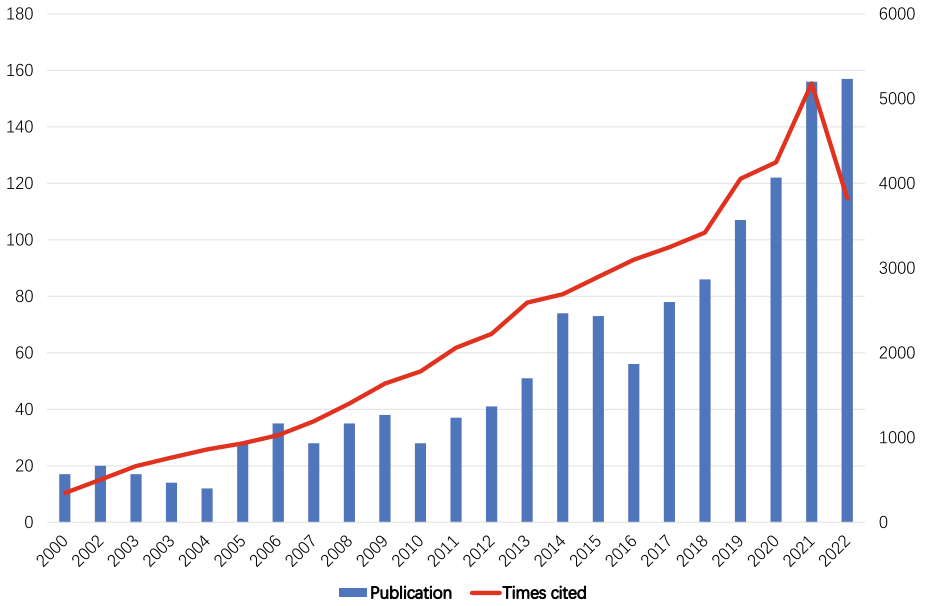


Fig. 11 Number of papers and citations for temporal action detection

rapid development, but the number of papers was still low. Compared with the field of action recognition, temporal action detection has great prospects in respect of development and application.

We also identified scholars from different regions and summarized them by country. As shown in Fig. 12, a global map is used to show the contributions of different countries

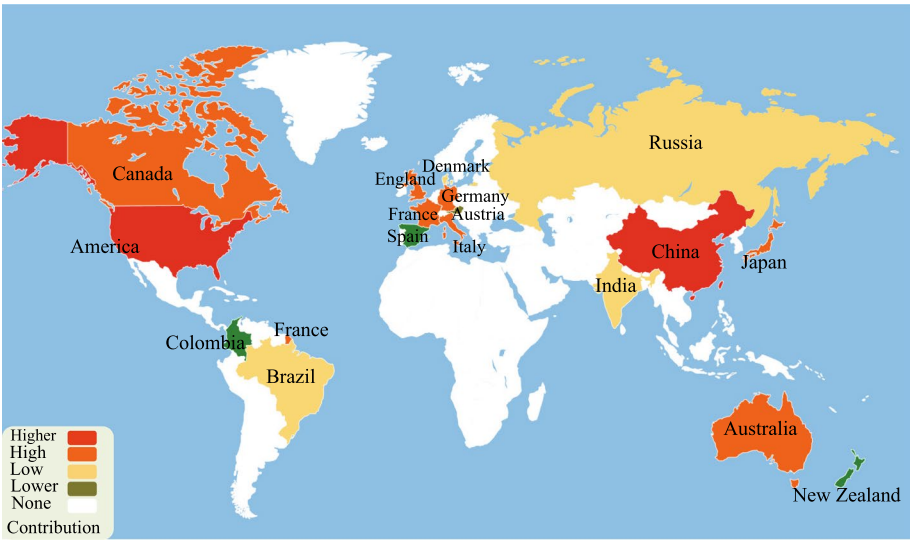


Fig. 12 Contributions of different countries to temporal action detection

through different shades of color. According to the color in the lower-left corner, we ranked the contribution degree from the highest to the lowest. The color red indicates that the number of papers contributed is more than 300; orange indicates that the number of papers contributed is between 200 and 300; gold indicates that the number of papers contributed is between 100 to 200; green indicates that the number of papers contributed is less than 100; white indicates that the relevant areas have not contributed to TAD research. It can be seen that China and the United States, followed by Japan, Canada, and some European countries, have made great contributions to temporal action detection research. More countries have begun to pay attention to research in respect of temporal action detection tasks.

4 Video feature extraction

Due to limited computer resources, a video cannot be directly applied to TAD as the input. Generally, the video needs to be input into a visual encoder. After processing by a visual encoder, the video can be represented by a series of visual features that are further processed for subsequent tasks. According to the history of video feature extraction, this section is divided into feature extraction via traditional methods and feature extraction via deep learning methods.

4.1 The traditional methods

Some early methods used manual features or local space-time descriptor operators as representations of videos to classify and detect video actions. Laptev (2005) proposed the space-time interest point (STIP) in 2003 by extending the Harris corner detector to 3D. SIFT and HOG were extended to SIFT 3D and HOG3D for action recognition by Scovanner et al. (2007) and Klaser et al. (2008), and others. Ke et al. proposed a cuboid feature (Ke et al. 2007) for behavior recognition in 2007. Sadanand and Corso established ActionBank (Sadanand and Corso 2012) for action recognition in 2012. The most representative algorithm is the dense trajectories (DT) algorithm proposed by Wang et al. in 2011. Firstly, the feature trajectories in the video frame sequence are obtained by the optical flow field, and then feature extraction is carried out based on the feature trajectories. However, feature extraction via the DT algorithm is often subject to environmental constraints. For this reason, Wang et al. (2013) proposed an improved dense trajectory (IDT) method in 2013. This is a more advanced video feature extraction method. The IDT descriptor shows how spatial and temporal signals can be processed differently. Instead of extending the Harris corner detector to 3D, it starts with densely sampled feature points in the video frame and uses the optical flow to track them. For each tracker corner, a different manual feature is extracted along the track. This method can weaken the influence of camera motion on feature extraction, making the IDT algorithm the best method with the best effect and stability before deep learning entered the field. However, this method requires much computation and has difficulty in dealing with large-scale datasets. Moreover, its features lack flexibility and extensibility.

4.2 Deep learning methods

Deep learning technology solves the problem of the feature extraction of large-scale datasets, enabling many datasets to be trained to extract the spatial and temporal features

of videos and generate a good model. At present, deep learning of human behavior recognition can be divided into three categories: CNN, RNN, and transformers (Hu et al. 2022c). Next, relevant work will be introduced in these three categories.

4.2.1 CNN feature extraction

Inspired by deep learning breakthroughs in the image field (Krizhevsky et al. 2017), various pre-trained convolutional networks (Jia et al. 2014) can be used to extract image features. Feature extraction methods based on convolutional neural networks can be divided into the two-stream CNN and the 3D CNN. In this section we review the two schemes.

(1) Two-stream 2D CNN

The progress of image recognition methods promotes two-stream convolutional networks (Hu et al. 2022a) for action recognition. Before this, many video action recognition methods were based on local spatiotemporal features of shallow high-dimensional coding. Simonyan and Zisserman (2014) first proposed a two-stream convolutional network containing space-time and optical flow in 2014. As shown in Fig. 13, the network consists of two parts that process data in both time and space dimensions; each network is made up of a CNN and the last layer is Softmax. Since two-stream convolutional networks only operate one frame (spatial network) or a single heap frame in short segments (temporal network), they have poor modeling ability for the time structure in a long time-range and limited ability to capture contextual relations.

In view of the shortcomings of the two-stream convolutional network, Wang et al. (Wang et al. 2016b) summarized two problems that need to be solved in 2016:

(a) How to design an effective video-level framework for learning video representation that captures long time structures;
(b) How to train a neural convolutional network model with limited samples.

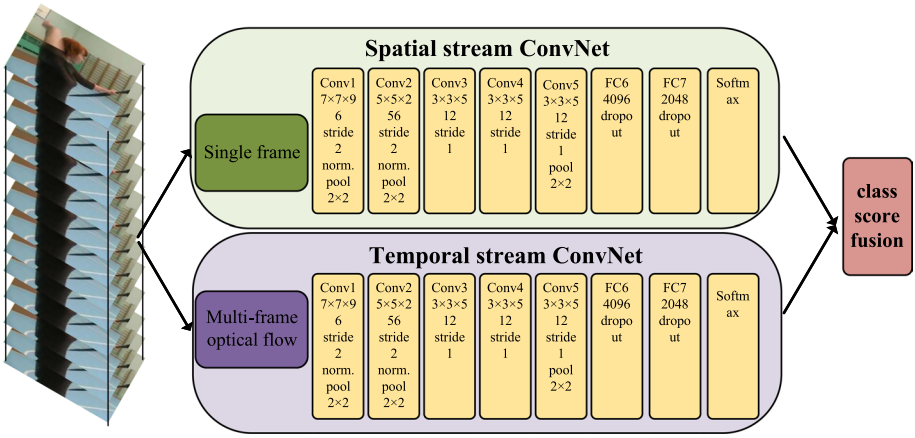


Fig. 13 Two-stream architecture for video classification. (Simonyan and Zisserman 2014)

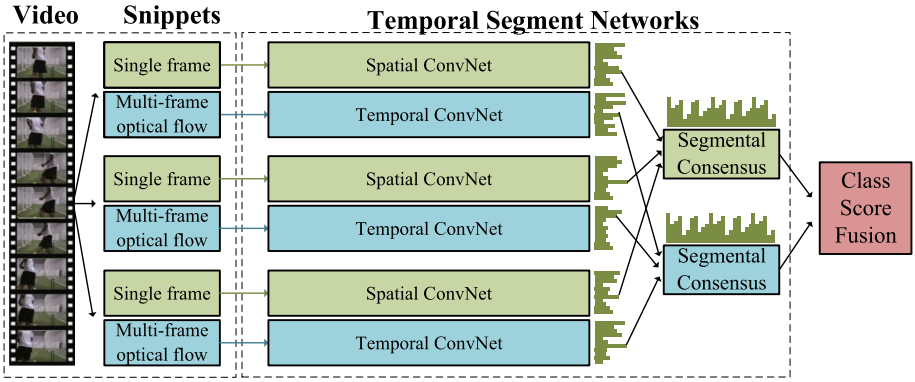


Fig. 14 TSN Module. (Wang et al. 2016b)

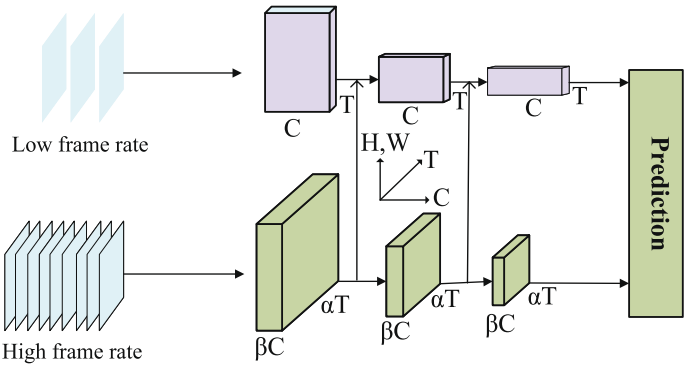


Fig. 15 SlowFast Model. (Feichtenhofer et al. 2019)

Therefore, Wang et al. proposed the classical TSN network and introduced a very deep neural convolutional network structure. The model can be combined with complete video information by using sparse time sampling and video-level supervision. As shown in Fig. 14, the TSN model framework is divided into spatial stream convolutional networks and optical stream convolutional networks. The processing objects are no longer single frames or single heap frames but sparsely sampled snippets. TSN then fuses the category scores of different segments through temporal segment networks. TSN enables end-to-end learning of long video sequences within a reasonable budget of time and computer resources.

In 2019, Feichtenhofer et al. (2019) proposed a slow-fast network inspired by the two-stream idea and biological research in respect of retinal ganglion cells in primate visual systems (Hubel and Wiesel 1965; Livingstone and Hubel 1988; Derrington and Lennie 1984; Felleman and Van Essen 1991; Van Essen and Gallant 1994), and achieved excellent performance. This model uses a slow path running at a low frame rate, which can be any convolution model (Tran et al. 2015; Feichtenhofer et al. 2017; Carreira and Zisserman 2017; Wang et al. 2018; Hu et al. 2023) and resolves static content in the video by capturing spatial semantics. There is also a fast path that runs at a high frame rate, capturing motion with good temporal resolution to analyze the dynamic content in the video. As

shown in Fig. 15, the two paths are the low frame rate and the high frame rate. T and C of the slow path are the benchmarks of the fast path. For video, the slow path samples T frames as the input. At the same time, the fast path needs to process high-frequency information; the whole process does not use the time domain downsampling layer, so the input is always -framed. The two paths are fused by a transverse connection and finally fed into a full connection layer for classification. The validity of the model has been proven using 6 datasets.

Video feature extraction based on deep learning has obvious performance advantages compared with traditional manual feature extraction methods. The two-stream network divides video sequences into time and space in a pioneering way, providing a research space for subsequent researchers. However, the speed of the two-stream convolutional network is slow, making it unsuitable for use with large-scale real-time video; 3DCNN can make up for this deficiency.

(2) 3DCNN

Another idea for video feature extraction is to expand the 2D convolution kernel used for image feature extraction into a 3D convolution kernel to train a new feature extraction network. The general approach of feature extraction algorithms based on 3D convolutional networks is to take a spatiotemporal cube formed by stacking a small number of continuous video frames as the model input. Then, the spatial and temporal representation of the video information is adaptively learned through the hierarchical training mechanism under the supervision of the given action category label.

In 2015, Tran et al. (2015) proposed a simple and effective spatiotemporal feature-learning method called C3D. Experiments show that 3D convolutional networks are more suitable for spatiotemporal feature learning than 2D convolutional networks and have significantly improved efficiency compared with two-stream methods. However, C3D is not learned from a full video. Therefore, the modeling ability for long-range spatiotemporal dependence is not strong. Diba et al. (2017) proposed time 3DCNN (T3D) in 2017, which strengthened the modeling of long-range spatiotemporal dependence. The time transition layer can model the depth of the time convolution kernel. T3D can efficiently capture short-, medium-, and long-term time information. Varol et al. (2017) proposed a long-term time convolution (LTC) network in 2017, which increased the time range of the 3D convolution layer at the cost of reducing spatial resolution and enhanced the modeling of the long-term time structure.

Due to the introduction of the time dimension, the network parameters become larger, and the training cost becomes increasingly higher. Some researchers aim to decompose 3D convolution. Qiu et al. (2017) proposed a pseudo-3D residual network in 2017. Pseudo-3D decomposes 3D convolution into a two-dimensional spatial convolution with a convolution kernel of 1×3×3 and a one-dimensional time convolution with a convolution kernel of 3×1×1 to simulate 3D convolution. Carreira et al. (2017) proposed the I3D model in 2017 using the same refinement idea. The main idea was to extend Inception's 2D model into a 3D model. I3D can achieve excellent performance after full pre-training in kinetics. Similarly, Tran et al. proposed a new spatiotemporal convolution module called R(2+1)D in 2018, which approximates the complete three-dimensional convolution with a two-dimensional convolution kernel and a one-dimensional convolution, thus separating the processing of space and time. Lin et al. (2019a) proposed the temporal shift module (TSM) network in 2019. TSM moves some features forward and backwards along the time

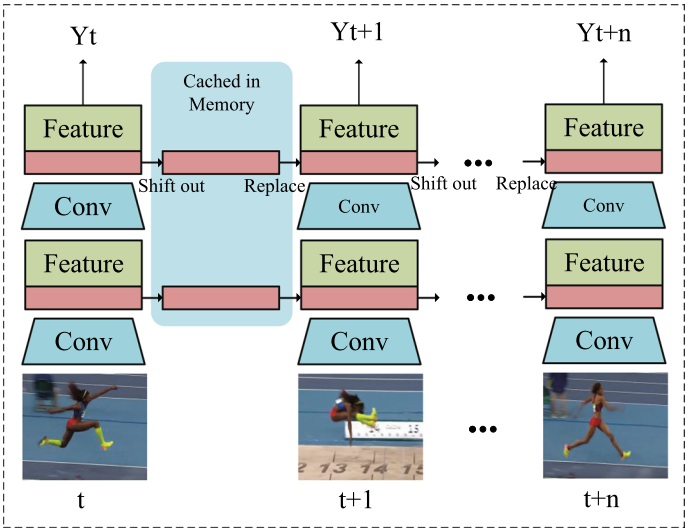


Fig. 16 TSM Model. (Lin et al. 2019a)

dimension, allowing the network to achieve the performance of a 3D CNN but maintain the complexity of a 2D CNN. As shown in Fig. 16, for the image input of a single frame, only the first 1/8 feature graphs of each residual block are saved and cached in memory during feature processing. The red box in the figure is the 1/8th feature of the cache. The author uses the cache feature map for the next frame to replace the first 1/8 of the current feature and 7/8 of the current feature map to create the next layer. As a result, TSM gives an inter-frame predictive delay that is almost identical to the 2D CNN baseline.

Compared with two-stream convolution, 3D convolution is faster and more efficient. However, the existing network cannot make full use of video temporal and spatial characteristics, and the recognition rate is low. Therefore, feature extraction methods for human action recognition still need to be optimized.

4.2.2 RNN feature extraction

RNNs can be used to analyze temporal data due to recursive joins in its hidden layer. However, traditional RNNs have the problem of disappearing gradients, and cannot effectively model long time series. Therefore, most current approaches adopt gated RNN architectures, such as LSTM (Hu et al. 2021), which can effectively model video-level temporal information.

Donahue et al. (2015) proposed the long-term recurrent convolutional network (LRCN) model in 2014. This model uses 2D-CNN as a feature extractor to extract frame-level RGB features and connects LSTM for prediction. In 2015, Yue-Hei Ng et al. (2015) proposed two methods for processing long videos. The first method explores various convolution time feature pool architectures, including conv pooling, late pooling, and slow pooling, and using the aggregation technology characterized by maximum pooling, frames can be extracted at a higher frame rate while still being able to extract the full video in the capture context. The second method uses a recursive neural network composed of LSTM units to model the video explicitly as an ordered sequence of frames. As shown in Fig. 17, the input

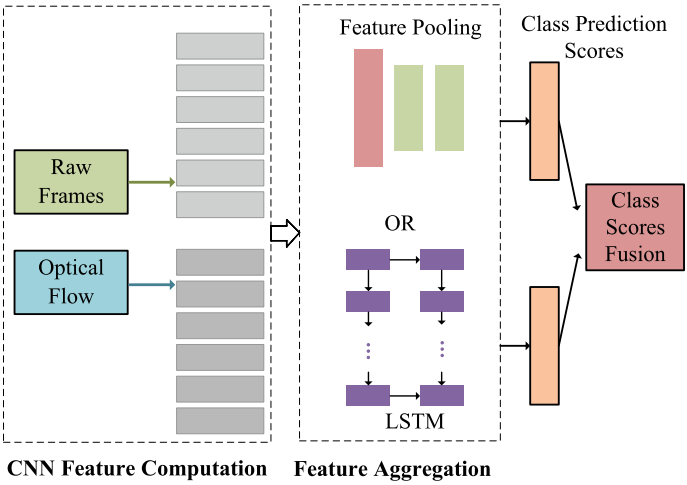


Fig. 17 Overview of NG'S approach. (Yue-Hei Ng et al. 2015)

original frame and optical flow enter the feature aggregation after CNN feature computation. LSTM networks run on frame-level CNN activations and can learn how to aggregate feature information over time. The network can share parameters over time, and both architectures are able to maintain a constant number of parameters while capturing a global description of the video's temporal evolution.

In the same year, Srivastava et al. (2015) proposed a recursive neural network based on LSTM, namely the encoder LSTM and the decoder LSTM. An encoder LSTM is used to map the input video to a fixed-length representation, and then the decoder LSTM decodes

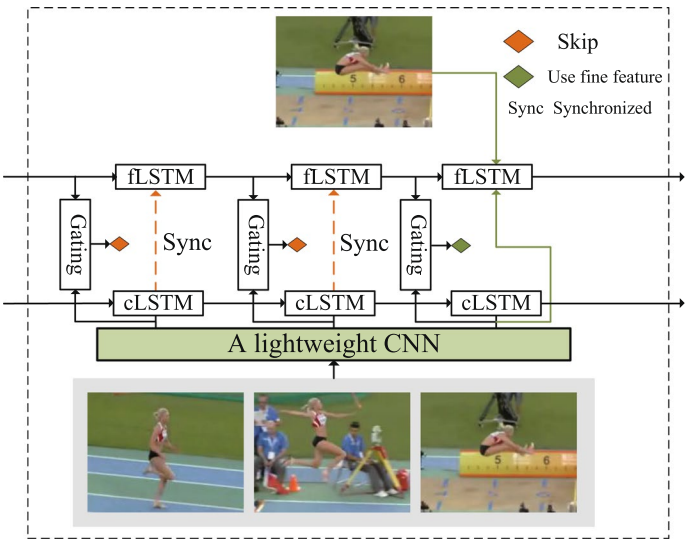


Fig. 18 Overview of the LITEEVAL approach. (Wu et al. 2019)

it to obtain the processed video features. In 2019, Wu et al. (2019) proposed a LITEEVAL framework based on rough and fine LSTM. As shown in Fig. 18, rough LSTM (cLSTM) is used to process the features extracted by lightweight CNN from the video and obtain rough features. It determines whether to calculate the fine features based on the rough features and historical information. If further checks are required, fine features are exported to update the fine LSTM (Flstm); otherwise, the two LSTMS are synchronized. Fine LSTM can obtain all the feature information it sees. Majd and Safabakhsh (2020) proposed a novel C^2 LSTM in 2020, which uses convolution and cross-correlation operators to learn the spatial and action features of videos and extract time dependence.

With the introduction of spatial and temporal attention, LSTM has undergone new development. Sharma et al. (Sharma et al. 2015) added spatial attention to the LSTM unit for the first time. The model recursively outputs the attention diagram and pays more attention to the spatial information of features. In 2019, Sudhakara et al. (Sudhakaran et al. 2019) introduced a recurrent unit with built-in spatial attention called long short-term attention (LSTA), which can spatially localize discriminative information on video input sequences. As shown in Fig. 19, LSTA extends the LSTM with two new components, the circular attention and the output pool. The first part (red) tracks the weight plot to focus on relevant features, while the second part (green) introduces high-capacity output gates. At the core of both is a pool operation ξ that enables smooth attention tracking and flexible output gating. To make full use of spatial features in the video, Li et al. (2018b) proposed the VideoLSTM model in 2016. VideoLSTM uses convolution and action-based attention mechanisms to obtain spatial correlation and action-based attention graphs in video frames. In 2021, Muhammad et al. (2021) proposed an attentional mechanism based on bidirectional long- and short-term memory, which includes a dilated convolution neural network (DCNN). DCNN extracts CNN features from input data, and the network can selectively focus on valid features in video input frames.

Video feature extraction methods based on 3D CNN usually perform spatiotemporal processing at limited intervals through window-based 3D convolution operations, in which each convolution operation only pays attention to the relatively short-term context of the video. At the same time, RNN-based methods recursively process video sequence elements, so they cannot model relatively long-term spatiotemporal dependence. However, transformers can be directly involved (Sun et al. 2022b).

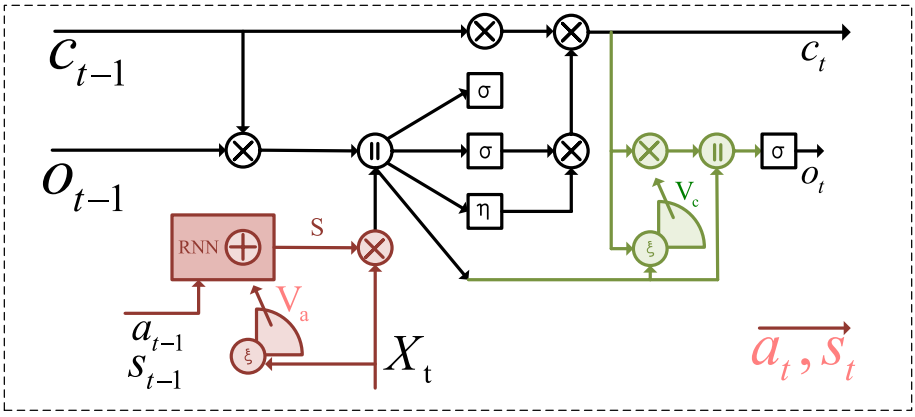


Fig. 19 Overview of the LSTA approach (Sudhakaran et al. 2019)

4.2.3 Transformer feature extraction

Transformer networks (Vaswani et al. 2017; Acheampong et al. 2021) have significant performance advantages and are becoming popular in deep learning. Compared with traditional deep learning networks such as convolutional neural networks and cyclic neural networks, transformers are more suitable for feature extraction because their network structure is easy to deepen and the model deviation is small (Ruan and Jin 2022); they perform well in long-term dependence modeling.

Given the success of transformers in natural language processing, more researchers are using transformers in video processing. Girdhar et al. (2019) proposed an action-based transformer model in 2019. This model uses the initial layer of space-time I3D to generate the basic features and then generates the boundary proposal using the regional proposal network (RPN). The basic feature map and each proposal are obtained through the action transformer to obtain the proposed features. An action transformer treats the feature graph for each particular topic as a query and the features from adjacent frames as keys and values. As shown in Fig. 20, the video footage is taken as the input, and the backbone network (usually the initial layer of I3D) is used to generate the spatiotemporal feature representation. The central frame of the feature map generates the bounding box proposal through the RPN, the feature map (filled with positional embedding) and each proposal obtains the proposal's characteristics through the "head" network. The head network is made up of action transformer units (Tx units) to generate the features to be categorized. QPr and FFN refer to the query preprocessor and feedforward network, respectively.

Bertasius et al. (2021) proposed the TimeSformer in 2021. TimeSformer is an adaptation of the standard transformer structure to video through learning spatiotemporal features directly from a series of frame-level patches (Dosovitskiy et al. 2020). TimeSformer removes the CNN entirely for the first time and applies temporal and spatial attention to each piece. Inspired by the vision transformer (ViT), Arnab et al. (2021) proposed a pure transformer model called ViViT in the same year. By degenerating different components of the transformer encoder in spatial and temporal dimensions, a large number of spatial-temporal markers encountered in the video can be effectively handled. As shown in Fig. 21, the model extracts space-time tokens from the input video and then encodes the space-temporal tokens through a series of transformation layers. Three effective model components are also shown in the figure, which handle the long sequence of tokens encountered in the video: factorized encoder, factorized self-attention, and factorized dot-product.

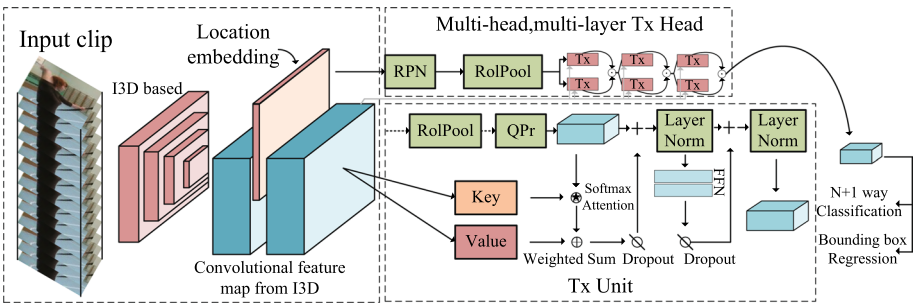


Fig. 20 Overview of action transformer (Girdhar et al. 2019)

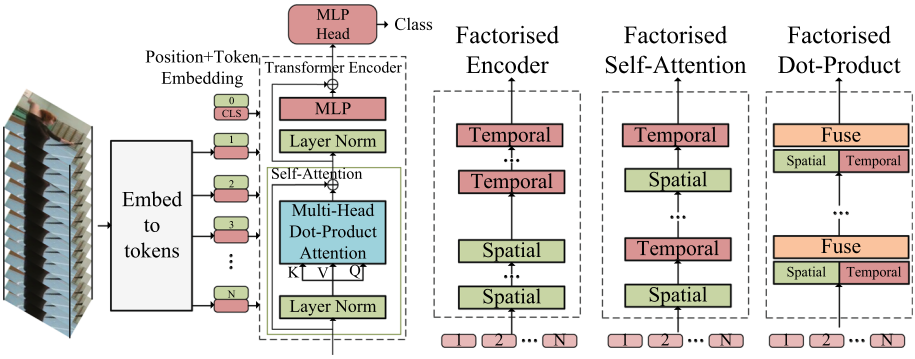


Fig. 21 Overview of ViViT (Arnab et al. 2021)

Only after pre-training a large amount of data can a pure transformer achieve better performance than a CNN, but it will inevitably require substantial memory and computing consumption. Zha et al. (2021) proposed a shifted chunk transformer with a pure self-attention block in 2021. In processing video frames, a pure transformer divides each frame into several local windows called image blocks and builds a layered image block converter. The converter uses locally sensitive hashing to enhance dot-product attention in each block (Hu et al. 2022c), thereby significantly reducing memory and computing consumption. In addition, to fully consider the action effect of the object, Zha et al. also designed a powerful self-attention module, namely the shifted self-attention module:the module explicitly extracts correlations from nearby frames. Furthermore, a frame-by-frame attention module clip encoder based on a pure transformer was designed to model the complex inter-frame relationship with minimal additional computational cost.

In addition,a series of Transformer-based models have recently emerged. For example, MotionFormer proposed by Patrick et al. (Patrick et al. 2021) in 2021 is used for video action recognition of people. This model proposes a new video Transformer framework called Trajectory Attention for modeling temporal correlation in dynamic scenes. As shown in Fig. 22, the figure shows the trajectory attention flow chart, which consists of two stages: the first step forms a set of ST trajectory markers for each space-time location st, and the second step utilizes 1D temporal attention operations Converge along these trajectories. In this way, trajectory attention can effectively accumulate information about the motion paths of objects in videos. In addition, MotionFormer also proposes a new method to calculate the quadratic dependence between memory and input size, which is especially important for high-resolution and long videos. However, MotionFormer has large calculation and memory overhead. Liu et al. (2021) introduced the Swin Transformer in the same year, addressing the challenges related to applying the Transformer from the language domain to the visual domain, such as the significant difference in the scale of visual entities and the high pixel resolution of images compared to words in text. As shown in Fig. 23, this illustration presents two consecutive Swin Transformer blocks, where W-MSA and SW-MSA have multi-head attention modules with regular and shifted window configurations, respectively. Swin Transformer introduces a hierarchical Transformer structure, utilizing shifted windows for computing representations. In various tasks, the experimental results have significantly surpassed previous state-of-the-art models, demonstrating the potential of Transformer-based models as visual backbone networks.

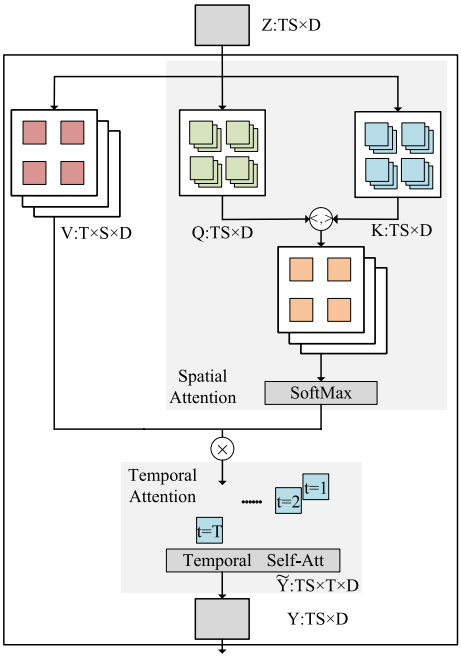


Fig. 22 Overview of motionformer (Patrick et al. 2021)

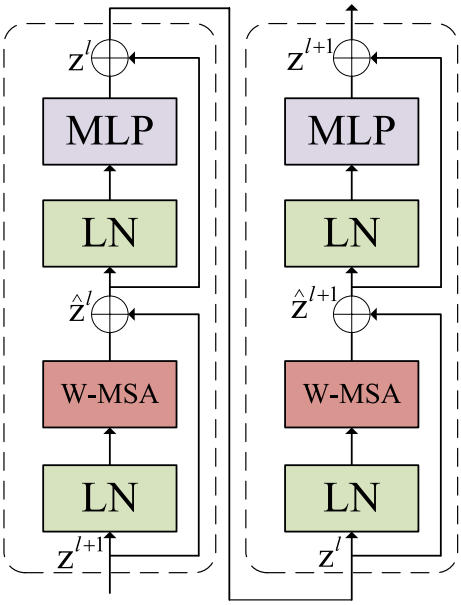


Fig. 23 Overview of swim transformer (Liu et al. 2021)

Although Swin Transformer has linear computational complexity,it may face issues such as memory limitations for particularly large image inputs. In 2021, Fan et al. (Fan et al. 2021) introduced the Multiscale Vision Transformers (MViT), a multi-scale visual Transformer model designed for video and image recognition that combines the concept

of multi-scale feature hierarchies with the Transformer model.MViT is better able to model the dense nature of visual signals through feature hierarchies at multiple scales. The MM-ViT proposed by Chen et al. (2022) in 2022 has been expanded and improved on the basis of MViT. MM-ViT incorporates multi-modal processing and cross-modal attention mechanisms,which handle data including motion vectors, residuals, and audio waveforms, and features three distinct cross-modal attention mechanisms.These cross-modal attention mechanisms can be seamlessly integrated into the Transformer architecture.

However,when a set of video frames is mixed in random order and is different from the original frame, it may be classified with the same label as the original recognition result; if this is the case, these models have clearly been over-fitted or biased toward other factors than the semantic information learned in respect of the actions. To solve this problem. Truong et al. (2022) proposed the DirecFormer model in 2022. The model learns the correct frame order in action videos by taking advantage of the direction of attention and the amount of attention between frames. Furthermore,to address the inability of conventional transformers to effectively quantify their forecast inaccuracies. Guo et al. (2022) proposed the uncertain guidance transformer (UGPT) in 2022, which treats the attention score of the transformer as a random variable to capture random dependencies and uncertainties in the input. The features extracted from the CNN are input into the UGPT after location encoding, and advanced embedding is the output.

Due to the large memory footprint of untrimmed video, the current advanced TAD is on top of the features of precomputed clip videos. These features may not be suitable for TAD. Specifically, the video encoder is trained to map different short films within an action sequence to similar outputs, thus predicting insensitivity to the time boundary of the action. As shown in Fig. 24, the image and video classification models can be fine-tuned to work with pre-training, thanks to the availability of large relevant datasets (such as ImageNet and UCF101 in the figure). However, the existing datasets for TAD tasks are too small for model pre-training or lack time boundaries, leading to low efficiency. Therefore, we believe

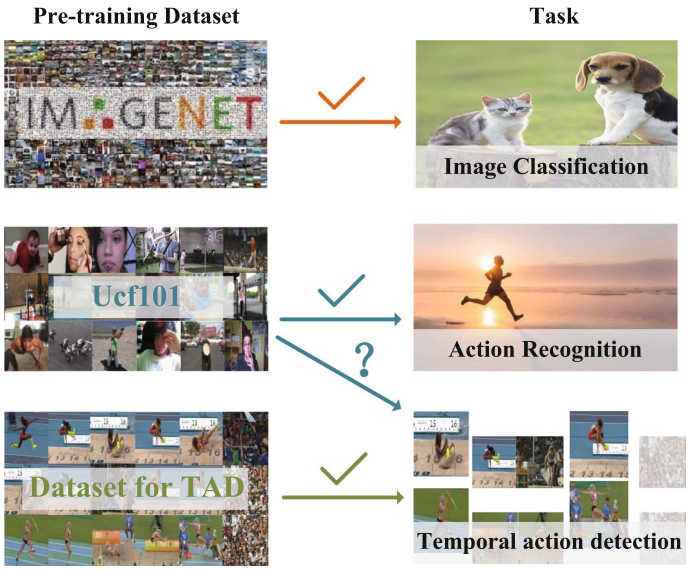


Fig. 24 Pre-training datasets for different tasks

that solving the limitations of the training design of TAD has great potential to improve the model's performance.

In 2021, Alwassel et al. (2021) proposed a novel supervised pre-training paradigm for editing. This paradigm not only trains the classification of foreground activities, but also considers background clips and global video information to improve time sensitivity. As large video datasets with time boundary annotations are difficult to collect, Xu et al. (2021a) designed the boundary-sensitive pretext (BSP) in 2021. They propose to transform the existing action classification dataset of clip videos to synthesize large-scale untrimmed videos with time boundary annotations. Specifically, they generate artificial time boundaries that are relatively consistent with changes in video content by splicing clips containing different classes, splicing two video lines of the same class, or manipulating the speed of different parts of the video instance. Xu et al. (2021b) proposed a simple and limited low-fidelity (LoFi) video encoder optimization method in the same year. They did this by introducing a strategy characterized by a new intermediate training phase, in which both the video encoder and the TAD head use lower temporal and spatial resolution (i.e., low-fidelity) for end-to-end optimization in small batch constructs. In 2022, Zhang et al. (2022) proposed a new unsupervised pretext learning method called pseudo action localization (PAL). PAL first builds training sets by cheaply converting existing large-scale TAD datasets. Then, two-time regions with random time lengths and proportions are randomly clipped from a video as pseudo actions. The model can align the pseudo-action features of the two synthesized videos.

Transformer addresses the problem that CNN- and RNN-based approaches cannot model relatively long-term spatial-temporal dependencies. The transformer can participate directly in the completion of video sequences through its extensible self-attention mechanism, thus effectively learning the remote spatiotemporal relationships in the video. The Table 2 summarizes the video feature extraction networks based on deep learning methods.

5 TAD according to the design method

Natural language processing and video understanding are different branches of artificial intelligence. The two are different in terms of application objects. The application object of natural language processing is two-dimensional text data, while the application object of TAD is three-dimensional action data. From the perspective of practical application, both are more in line with the characteristics of unknown information and rich information in real scenes, so there is a certain correlation in processing ideas and methods. Natural language processing can be seen everywhere in the design of TAD.

Natural language processing presents complex and challenging tasks related to languages, such as machine translation, questions and answers, and summaries. Nonlinear programming involves designing and implementing models, systems, and algorithms to solve the practical problems of understanding human language (Lauriola et al. 2022). Thanks to recent advances in deep learning, the performance of natural language processing applications has been improved in an unprecedented way, attracting increasing interest from the machine learning community (Kotsiantis et al. 2006). For example, the most (Wang et al. 2021a) advanced phrase-based statistical methods in machine translation have been gradually replaced by neural machine translation (Yadav and Vishwakarma 2020). Neural machine translation involves large deep neural networks that achieve better performance (Bahdanau et al. 2014). After the advent of text vectors and unsupervised pre-training, the

Table 2 Performance of deep learning methods in UCF101, HMDB-51 and Kinetics-400

	Methods	Time	Input	Accuracy rate(%)			Code resources
				UCF-101	HMDB-51	Kinetics-400	
CNN	Two-stream Simonyan and Zisserman (2014)	2014	RGB, Flow	88.0	59.4	–	https://github.com/jeffreyyihuang/two-stream-action-recognition
	C3D Tran et al. (2015)	2015	RGB	90.4	–	–	https://github.com/facebookarchive/C3D
	TSN Wang et al. (2016b)	2016	RGB, Flow	94.2	69.4	–	https://github.com/yjxiong/tsn-pytorch
	I3D Carreira and Zisserman (2017)	2017	RGB	97.9	80.2	–	https://github.com/piergiaj/pytorch-i3d
	P3D Qiu et al. 2017)	2014	RGB	93.7	-	–	https://github.com/qijiezhao/pseudo-3d-pytorch
	R(2+1)DTran et al. (2018)	2018	RGB	97.3	80.7	–	https://github.com/jfzhang95/pytorch-video-recognition
	TSM Lin et al. (2019a)	2019	RGB	95.9	73.5	–	https://github.com/ucas010/temporal-shift-module
	Slow-FastFeichtenhofer et al. (2019)	2019	RGB	-	–	79.8	https://github.com/facebookresearch/SlowFast
RNN	LRCN Donahue et al. 2015)	2015	RGB,Flow	88.0	59.4	–	-
	Ng et al. Yue-Hei Ng et al. (2015)	2015	RGB,Flow	88.6	–	–	https://github.com/face book/C3D
	Srivastava et al.Srivastava et al. (2015)	2015	RGB,Flow	75.5	44.1	–	-
	VideoLSTM Li et al. (2018b)	2016	RGB,Flow	79.9	43.3	–	https://github.com/zhenyangli/VideoLSTM
	Sharma et al.(2015)	2016	RGB,Flow	84.9	71.3	–	–
	C^2LSTM Majd and Safabakhsh (2020)	2020	RGB	92.8	67.1	–	–
	BiLSTM Muhammad et al. (2021)	2021	RGB	98.3	80.0	–	–

Table 2 (continued)

	Methods	Time	Input	Accuracy rate(%)			Code resources
				UCF-101	HMDB-51	Kinetics-400	
Transformer	Girdhar et al.(2019)	2019	RGB	–	–	–	https://github.com/ppriyank/Video-Action-Transformer-Network-Pytorch-
	TimeSformer Bertasius et al. (2021)	2021	RGB	–	–	80.7	https://github.com/facebookresearch/TimeSformer
	ViViT Arnab et al. (2021)	2021	RGB	–	–	84.9	https://github.com/rishikksh20/ViViT-pytorch
	MViT Fan et al. (2021)	2021	RGB	–	–	81.2	https://github.com/jeffreyyihuang/two-stream-action-recognition
	MotionFormer Patrick et al. (2021)	2021	RGB	–	–	81.1	https://github.com/facebookresearch/Motionformer
	Swin Transformer Liu et al. (2021)	2021	RGB	–	–	–	https://github.com/microsoft/Swin-Transformer
	MM-ViT Chen and Ho (2022)	2022	RGB	98.8	–	–	–
	DirecFormer Truong et al. (2022)	2022	RGB	–	–	82.8	–
	UGPT Guo et al. (2022)	2022	RGB	–	–	–	–

last real boost in NPL was the transformer model (Vaswani et al. 2017). The most popular pre-trained transformer model is BERT (Devlin et al. 2018). BERT aims to pre-train deep bidirectional representations in the unlabeled text by jointly moderating left and right contexts in all layers. Inspired by BERT, several pre-training models followed, such as RoBERTa (Liu et al. 2019b), ALBERT (Lan et al. 2019) and DistilBERT (Sanh et al. 2019). Other related approaches based on the same concept are generation pre-training (GPT) (Radford et al. 2018, 2019), Transformer XL (Dai et al. 2019), and its extension XLNet. Today, these methods continue to achieve excellent performance for a wide range of natural language processing tasks, such as question and answer (Garg et al. 2020; Shao et al. 2019; Kumar et al. 2019), text classification (Sun et al. 2019), sentiment analysis (Abdelgwad 2021), biomedical text mining (Lee et al. 2020a), and named entity recognition (Yu et al. 2019).

Temporal action detection can be regarded as the time version of image detection. The research in respect of TAD relies heavily on the timing proposal effect of the target action, and the video data have a complicated structure and different durations of action. In 2017, ActivityNet, a video understanding competition, proposed the concept of the TAD task, which introduced the problems of locating multiple target action video surveillance analyses, network video retrieval, and other tasks. According to the design method, TAD is divided into three categories: anchor-based, boundary-based, and query-based. Anchor-based TAD (5.1) generates a time proposal by assigning dense and multiscale intervals with predefined lengths to evenly distributed time positions in the input video. Boundary-based TAD (5.2) does not set predefined proposals, so proposals with precise boundaries and flexible durations can be generated. Query-based TAD (5.3) proposes to map a set of learnable embeddings to action instances to generate action proposals directly. In this section we follow the three design methods (Fig. 25).

5.1 Anchor-based methods

Anchor-based approaches are also known as top-down approaches. The designer first designs multiscale anchoring on each network of the feature sequence. Then, action classification and boundary regression are carried out according to the candidate proposals. The methods are mainly based on using sliding windows or anchor kernels to generate temporal action proposals. This section introduces anchor-based models in detail, from the earliest to the latest.

Most previous methods have relied on hand-selected features with significant performance improvements. Some researchers have attempted to combine IDT (Wang and Schmid 2013) with the appearance features self-extracted by frame-level deep networks (Oneata et al. 2014b). Due to the great success of deep learning in object detection, Shou

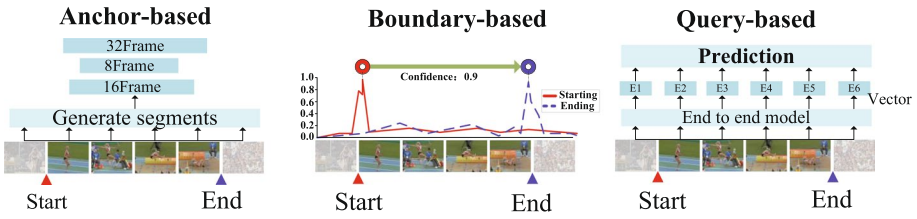


Fig. 25 The three design methods are Anchor-based,Boundary-based,Query-based (Vahdani and Tian 2022)

et al. (Shou et al. 2016) first proposed the S-CNN model in 2016 by means of region-based convolutional neural networks (R-CNNs) (Girshick et al. 2014) and their modifiers. The S-CNN method uses fixed windows of multiple sizes to process video clips and then uses a three-stage S-CNN for processing. As shown in Fig. 26, the overall framework is divided into three parts:

(a) Step 1 is called the proposal network. The proposal generation network is used to calculate the probability of action in all video clips;
(b) Step 2 is called the classification network. Classification networks are used to classify different actions and backgrounds;
(c) Step 3 is called the localization network. The similar output of the location and classification networks is still the probability of each action.

In training, overlap loss based on the IOU score is increased to make better use of the overlap rate. In theory, this method only has a high degree of overlap, the better the effect, but it produces a lot of redundancy. In addition, the convolution kernel of standard C3D used by SCNN is 3, and the receptive field is too small, so only short time-sequence information can be used.

Considering the limitations of the sliding window method, Gao et al. (2017a) proposed in 2017 that the TURN model could reduce the amount of computation and improve accuracy. The main idea of this model is to draw on the boundary regression in Faster-RCNN (Ren et al. 2015). As shown in Fig. 27, the video should first be divided into fixed-size units, such as 16 video frames, and each group should learn one feature (using C3D). Then, each group or multiple groups will be used as the central anchor unit (referring to Faster-RCNN) and expand to both ends to create a fragment pyramid. Next, coordinate regression is performed at the unit level, and the regression model has two sibling outputs. The first output is the trustworthiness score, which is used to determine whether the action is present in the fragment. The second output is the time coordinate regression offset. Compared with frame-level coordinate regression, unit-level coordinate regression is easier to learn and more efficient. An innovative new architecture based on time-coordinate regression and capable of running at over 800fps is proposed.

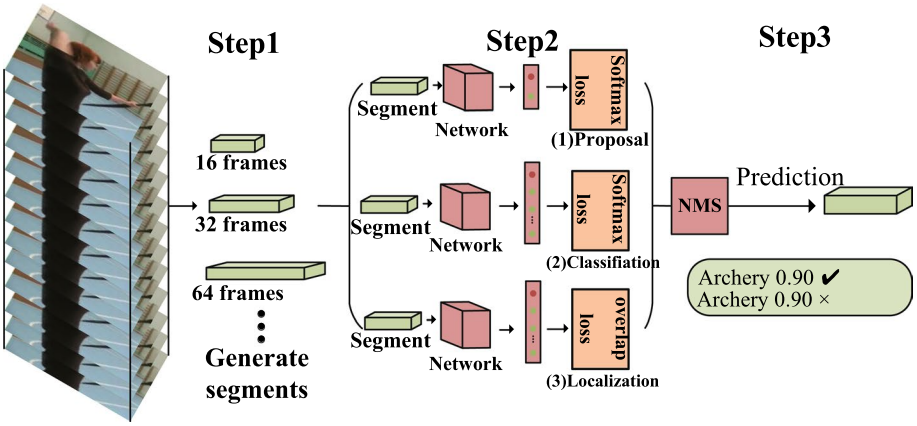


Fig. 26 Segment-CNN Model (Shou et al. 2016)

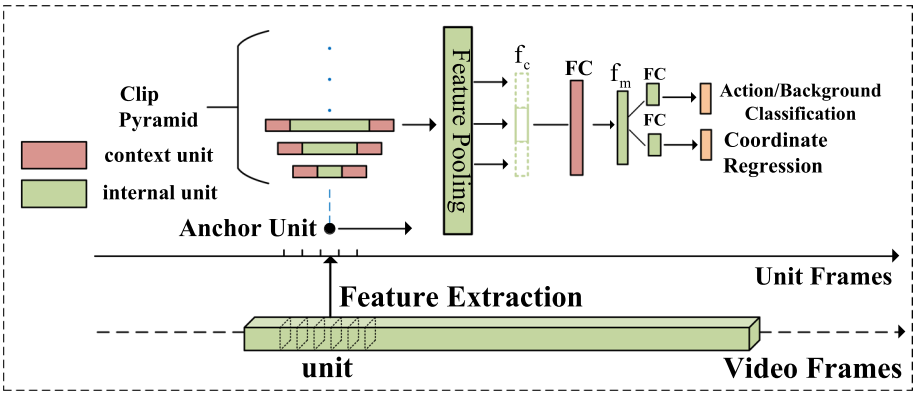


Fig. 27 Architecture of TURN (Gao et al. 2017a)

DAPs (Escorcia et al. 2016) and sparse prop (Heilbron et al. 2016) use average recall and the average number of search proposals (AR-AN) to evaluate TAP performance. There are two problems with the AR-AN metric:

(a) the correlation between the AR-AN of TAP and the mean average precision (mAP) of action positioning has not been discussed
(b) the average number of proposals retrieved correlates with the average video length of the test dataset, which makes AR-AN less reliable for evaluation across different datasets. Spatiotemporal action detection (Yu and Yuan 2015; Wang et al. 2016a) uses the recall and proposal number (R-N), but this index does not consider the video length.

The TURN method adopts the brand-new temporal action proposal (TAP) indicator AR-F to solve the above two problems. However, it does not fundamentally solve the problem of inaccurate division of the action boundary.

Because most current models obtain good results through predetermined anchor points, they are susceptible to the interference of a large number of outputs and different anchor sizes. Lin et al. (2021) proposed a completely anchor-free temporal positioning method in 2021, which is more portable. The overall framework is divided into three parts:

(a) In the video feature extraction part, an I3D network is used to extract features and finally transform them into a 1D feature pyramid;
(b) Rough prediction, which is the first part of the anchoring, predicts the length of each video segment for each pyramid layer and classifies that segment;
(c) In fine prediction, significant boundary features are found for proposals generated in rough prediction. Then, in turn, the proposed boundary is optimized with the boundary features to obtain fine prediction results and output the confidence of the proposal.

As shown in Fig. 28, given a video as input, the I3D model is used to extract features and construct 1D time pyramid features. Next, each pyramid feature is fed into two basic prediction modules: a regressor generating a rough boundary score and a classifier generating a rough category score. Finally, the saliency-based refinement module adjusts class scores and the start and end boundaries, and predicts the corresponding quality scores for each

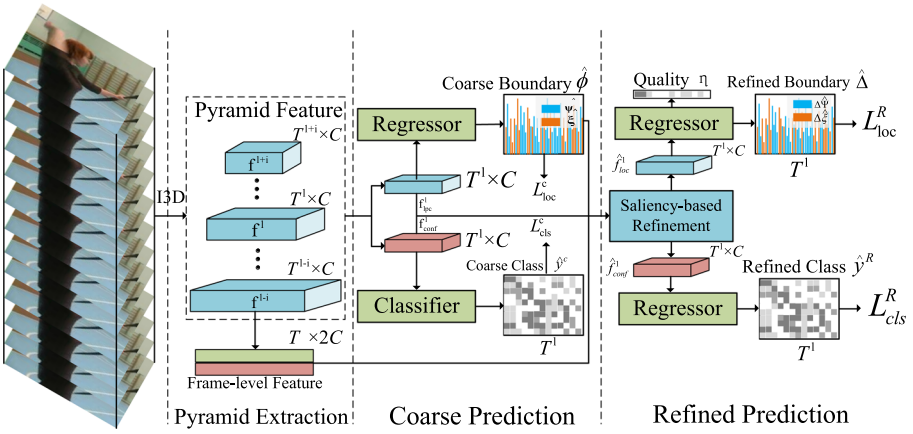


Fig. 28 Architecture of AFSD (Lin et al. 2021)

rough proposal. The main contribution of this method is its use of fewer parameters and outputs and obtaining a good performance. The influence of boundary features is explained by making full use of an unanchored frame. Finally, the method proposes a new consistent learning strategy for a better learning boundary control model. Yang et al. (Yang et al. 2020) showed that although the results of anchor-free methods are weak, there is evidence in the task of target detection (Girshick et al. 2014) that such methods should, in principle, be compared with anchor-based methods.

Given the great success of target detection in images, researchers have begun to apply Faster R-CNN to video TAD. However, several challenges have arisen in shifting the field, and the first is how to deal with the dramatic change in the duration of action, from a still image to a moving video; the second is how to use timing information. The moments before and after the action instance contain key information for positioning and classification, which is to some extent more important than spatial context information. The final issue is how to fuse multi-stream features. To address these challenges, Chao et al. (2018) proposed the TAL-Net network in 2018. TAL-Net uses a multi-tower network and dilatation time convolution to strengthen alignment between the receptive field and the span of the anchor. It uses a multiscale architecture to alter the receptive field so that it can adapt to continuous changes in action duration. By expanding the receptive field, TAL-Net can make better use of the time background to generate candidate proposals and classification. As for the feature fusion method, TAL-Net proposes a two-stream frame late-fusion scheme. Conceptually, this is equivalent to performing traditional late fusion in the proposal generation and action classification phases. The experimental results prove the feasibility of feature fusion in the later stage.

The above method can also be called the two-stage method. The two-stage method first generates the action candidate proposal, and then the generated candidate proposal is classified in the second stage. Most of the current models are inspired by R-CNN. The model shown in (a) in Fig. 29 comprises a time candidate window with a high recall rate provided by the anchor core introduced above, and then the candidate proposal is passed to the later classification stage. The two-stage approach treats the proposal and classification as two separate sequential processing stages, which inhibits collaboration between them and leads to double counting between the two stages. Figure 29b shows an end-to-end trainable

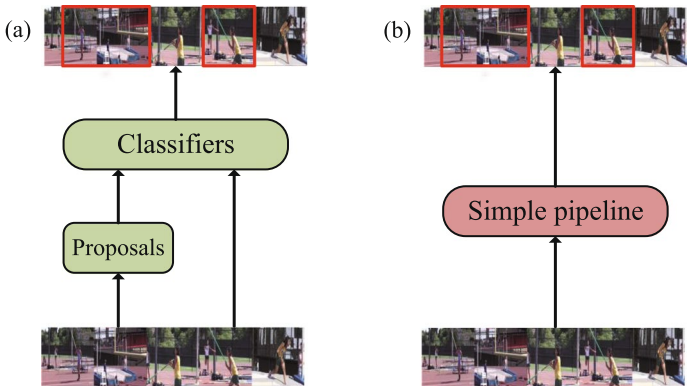


Fig. 29 **a** two-stage method, **b** one-stage method

approach that tightly integrates proposal generation and classification to provide a more efficient architecture for uniform TAD. Currently, most of the one-stage methods are based on the predefined anchor core, and the anchor-related hyperparameters need to be set with the knowledge of the action distribution in advance.

The single-stage model synchronizes proposal-making and proposal classification. Currently, most models adopt the two-stage model, which first proposes and then classifies the proposal. One inevitable problem with such a framework is that the boundaries of the action examples are already defined in the classification step. To solve this problem, in 2017 Lin et al. (2017) proposed a single-shot temporal action detection network (SSAD) based on a one-dimensional time convolution layer. SSAD can directly detect action instances in uncut videos and skip the generation of proposals. In SSAD, features of the single-lens object detection models SSD (Liu et al. 2016) and YOLO (Redmon et al. 2016) are adopted. As shown in Fig. 30, the model includes three modules:

(a) Base layer: The input video feature sequence is processed, the feature length is shortened, and the receptive domain is expanded;
(b) Anchor layer: This uses time convolution to reduce the feature map and output the anchored feature map;
(c) Prediction layer: The class, confidence, and position of each action instance are obtained by anchoring the feature graph. SSAD's biggest contribution is to eliminate the candidate proposal-generation step and implement an end-to-end framework.

Buch et al. (2019) proposed the single-stream temporal action detection (SS-TAD) network in the same year. Like SSAD, it is a single-emitter detector, and the design inspiration came from the single-emitter object detectors YOLO and SSD. The model consists of two parts. Input visual coding is used to encode low-level space-time information for the video (similar to how SSAD uses the C3D model to extract low-level space-time information). The two recurrent memory modules can effectively converge the context information to integrate it with the TAD task and output the final action instance's time boundary and confidence score. SS-TAD uses dynamic semantic constraints in the semantic subtasks of TAD to improve training and testing performance. Both SSAD and SS-TAD are end-to-end network models and reduce the number of times required to input video streams. The

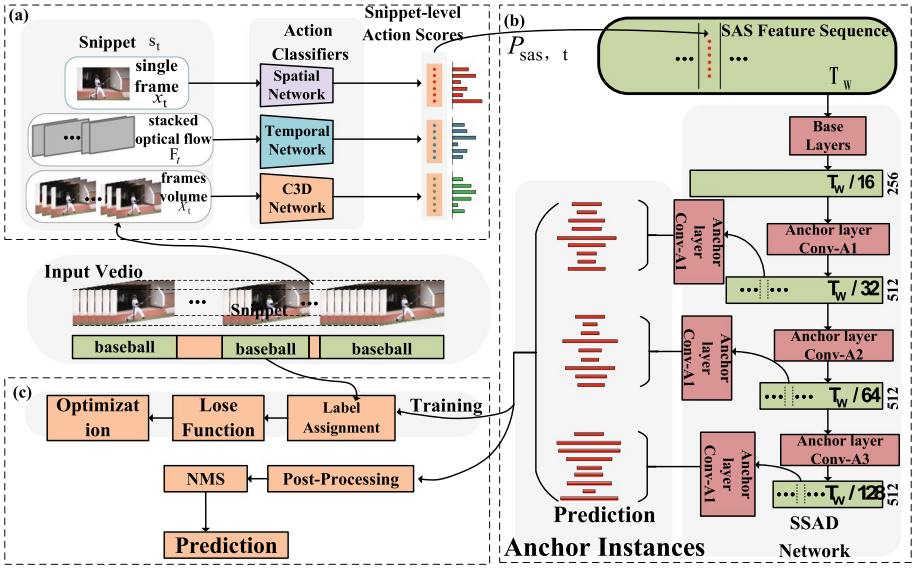


Fig. 30 The framework of the SSAD network (Lin et al. 2017)

feature encoding from the two memory modules is used to output the final time bound and associated class scores for the final output detection. As shown in Fig. 31, the SS-TAD model takes a video stream as input and then represents each non-overlapping "time step" t with the visual code in the δ frame. This visual coding acts as input to the two recurrent memory modules, making both modules semantically constrained to learn proposals and classifier-based features. These features are combined before providing the final TAD output. In contrast to previous work, the authors' approach provides end-to-end temporal action detection through a single pass of the input video stream.

Currently, most TAD models are designed with reference to image detection. Due to the fixed timescale, there may be a problem with robustness. In addition, the ability to detect complex actions is weak. Long et al. (2019) proposed Gaussian temporal awareness networks (GTAN) in 2019. The core innovation of GTAN is the use of a set of Gaussian kernels to simulate the time structure and optimize each generated action proposal. The addition of Gaussian kernels can represent action proposals of different sizes, and

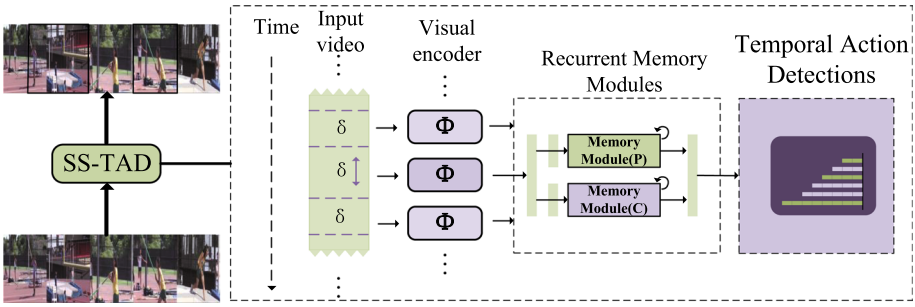


Fig. 31 SS-TAD model architecture (Buch et al. 2019)

the corresponding Gaussian curves can show the context of the generated action proposals. Specifically, a GTAN uses two convolution layers and the maximum pooling layer to shorten the feature mapping and increase the receptive field's time. Then, a series of one-dimensional time convolution layers (anchor layers) continuously shortens the feature map and outputs an anchoring feature map composed of the features of each cell (anchor layers). At the top of each anchor layer, Gaussian kernels are learned for each cell to dynamically predict a specific interval of action proposals corresponding to that cell. In addition, inspired by the idea of time action grouping in (Zhao et al. 2017), a hybrid convolution is included that can mix multiple Gaussian kernels to capture an action proposal of any length. GTAN offer improved performance compared to both two-stage and single-stage technologies.

Most of the proposed TAD models generate candidate proposals by subdividing them into frames or cutting them into video fragments. These methods are limited to local video information and cannot take advantage of video context relations. Based on traditional detection and linking strategies, in 2020 Li et al. (2020) proposed a single-stage model that can be trained end-to-end and a new coarse-to-fine action detector (CFAD). The CFAD uses two space-time action tubes, thick and thin. The strategy first estimates the thicker action tubes and then selectively refines these action tubes at critical points in time. An action tube is generated through the rough module and the refinement module. The rough module is designed to solve the problems of a lack of global information and inefficiency in previous detection and linking modes. In a global sense, it uses complete tube shape information to oversee tube regression. In addition, a parametric modeling scheme is introduced to describe the action tube. Instead of predicting a large number of box positions per frame, the rough module predicts only a few trajectory parameters to describe tubes of different durability. As a result, the module learns a robust notation that accurately and efficiently describes changes in the action tube. The refinement module delves into each pipe's local environment to find critical time locations to improve the estimated action pipe further, thereby improving overall inspection performance and efficiency. To correctly refine the action tube, CFAD includes a labeling algorithm to generate labels that can guide the learning of key timestamp selection. As shown in Fig. 32, the video frame is sent to the temporal proposal network (TPN) through a 3D CNN. A rough result is created in the rough module, and then pass the key time points in the video clip (the green point in the figure) into the fine module for processing so as to find the precise time position. "D Conv Head" means cascading NL-3D ("NL" denotes a non-local block (Wang et al. 2018)). The 2D Conv head block represents a cascading 2D spatial convolution. CFAD can achieve this 3.3 times faster than its nearest competitor.

The mainstream one-stage methods still rely on anchoring to generate candidate proposals; they lack generalization ability and cannot fully reflect the performance in complex, changeable action videos. In 2021, Ning et al. (2021) proposed a TAD called the selective receptive field network (SRF-Net), which is similar to AFSD's two-stage model and was designed to eliminate the anchoring method. SRF-Net can directly estimate the offsets and action categories at each point in time in the feature graph. Inspired by the selection mechanism in SKNet (Li et al. 2019), Ning et al. studied the selection mechanism of the receptive field in depth and creatively proposed the building block of selective receptive field convolution (SRFC), which can automatically adjust the size of the receptive field. The SRFC module can adaptively adjust the size of its receptive field according to multiple scales of the input information of each time position in the feature map. The different receptive fields at each time location are more effective on a specific timescale, which results in the classification and regression of corresponding action suggestions. FCOS

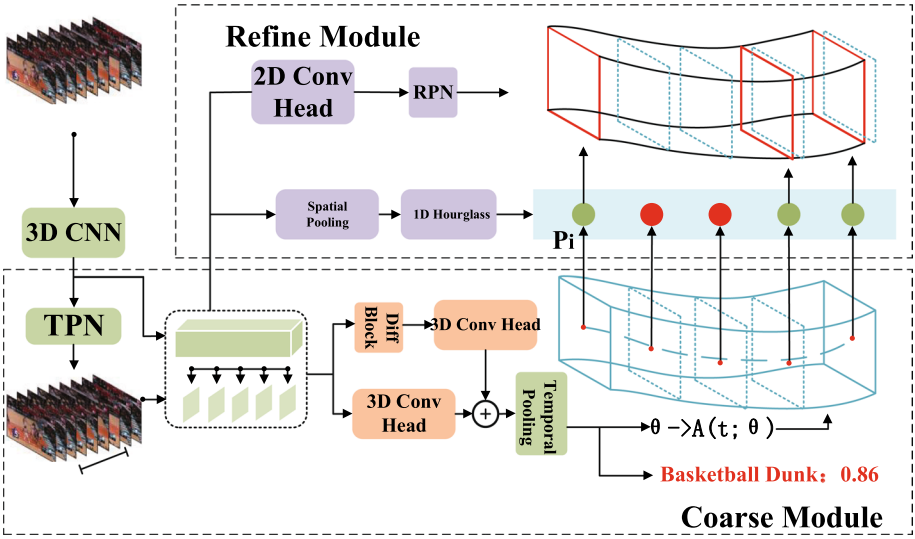


Fig. 32 Overview of the CFAD framework (Li et al. 2020)

(Tian et al. 2019) was referenced in the design of the prediction head, which uses a similar strategy to suppress some low-quality action proposals generated far from the center of the action instance. The SRFC block function is implemented in three steps:

(a) Split: the original feature map is split to generate three extendable time feature maps and cover the whole time-range through expansion;
(b) Fusion: this controls the flow of information in the three time-feature graphs through the attention mechanism;
(c) Selection: this uses cross-channel soft attention mechanisms to select different information scales. SRFC undoubtedly provides a new idea in terms of an anchor-free approach.

Time representation is the cornerstone of modern action detection technology. Most of the current advanced methods rely on the dense anchoring scheme, in which the anchoring uses a discrete network to sample uniformly in the time domain and then regress the exact boundary. In 2022, Wang et al. (2022a) proposed recurrent continuous localization (RCL), which learns a completely continuous anchoring representation and is an explicit model conditioned on video embedding and time coordinates. Specifically, the model is based on an explicit model conditioned on video embedding and time coordinates, ensuring the ability to detect segments of arbitrary length. The key idea is to use deep neural networks to directly regress confidence scores from successive anchor points. Thus, precise line segments can be extracted by searching for local maxima in a continuous function. This method uses the concept of continuous anchored representation to achieve high-fidelity action detection. Unlike ordinary anchor-based detection techniques, this technique discretizes fragments into regular grids for measurement (Lin et al. 2019b), generating estimates in a continuous field. The proposed continuous representation can be intuitively understood as a learning location condition classifier whose confidence score is determined by video features and time coordinates. As shown in the

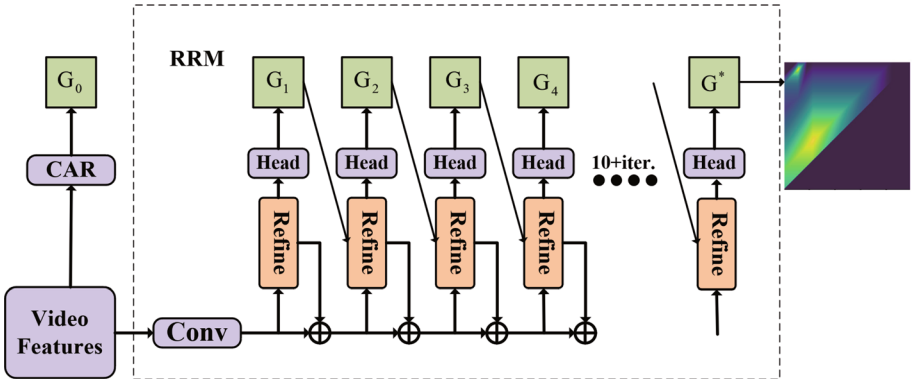


Fig. 33 RCL model. (Wang et al. 2022a)

Fig. 33, a feature encoder is first used to extract temporal video features, which are then fed to a continuous anchored representation (CAR) for predicting continuous confidence graphs with scaled invariant sampling strategies. Finally, a recursive refinement module (RRM) is entered to update the confidence graph by iteratively refining the uncertain region.

The above is the most representative anchor-based network model. S-CNN (Shou et al. 2016) has a fixed sliding window. In principle, good results can be achieved as long as the overlap rate is high enough, but redundancy will be generated correspondingly. Therefore, to improve the accuracy and reduce the number of parameters, TURN (Gao et al. 2017a) and TAL-Net (Chao et al. 2018) used the idea of an R-CNN and adopted boundary regression to generate candidate proposals. However, (Shou et al. 2016; Gao et al. 2017a; Chao et al. 2018) are all two-stage models, which inevitably have the following problems:

(a) Proposal generation and classification are trained separately, but in reality, it is preferred that they be trained together;
(b) The two-stage approach takes more time;
(c) Using the sliding window method results in similar time boundaries for action instances due to finite-sized windows.

Conversely, the single-stage method can generate and classify action candidate proposals in one instance, which results in true end-to-end learning. SSAD (Lin et al. 2017) and SS-TAD (Buch et al. 2019) learn from the single object detectors YOLO and SSD to extract low-level space-time information. In addition to using YOLO and SSD, the design innovation of single-stage methods like GTAN (Long et al. 2019) uses a Gaussian kernel to optimize the timescale to simulate action proposals of various sizes. CFAD (Li et al. 2020) uses two coarse refinement modules to improve the efficiency of generating accurate action tubes and remove traditional detection and linking strategies. However, anchoring representations can only provide rough recommendations (Li et al. 2020; Wang et al. 2022a) and are all aimed at obtaining a more continuous, fine-grained anchoring representation, and providing accurate estimates of the target region. These methods exhibit excellent performance and are capable of handling large duration periods. Instead of representing full segments, (Lin et al. 2021; Ning et al. 2021) employ an anchor-free approach, using a central point to represent a direct regression of the start and end times.

Anchor-based methods have introduced the idea of R-CNNs in target detection in a pioneering way and achieved good results. However, because the duration of live ground action instances varies greatly from video to video, these methods require a lot of computation when placing dense candidate proposals. In addition, the internal context of the scheme adopted in an anchor-based approach can obtain reliable confidence scores but cannot generate accurate boundaries. The Table 3 provides a summary of anchor-based methods, in which the feature types used by different network models and corresponding improvements and code websites are given.

5.2 Boundary-based methods

Boundary-based methods deal with imprecise boundary problems and directly evaluate each pair of matches in a video sequence. They first predict the boundary confidence of all frames and then use a bottom-up grouping strategy to match the start and end pairs. These methods extract boundary information from local windows and models using local context. They forgo the regression process and directly generate confidence scores for densely distributed proposals. This section introduces boundary-based models in detail.

Since the length of the untrimmed video is random, a method that relies on sliding windows will encounter difficulties in adapting to actions of different lengths. As a result, a large number of windows with different scales and a small sliding step size are needed, which causes an increase in computing costs. In addition, the capture of a complete action and the fuzzy distinction between actions make it difficult to accurately locate the start and end points. Xiong et al. (2017) proposed a model in 2017 that could be applied to uncut videos of different lengths and could accurately locate the time boundary of the action. The model consists of two parts: the generation of temporal proposal candidates and the classification of generation candidates. The TAG model is proposed in the temporal proposal candidate generation stage. Unlike the sliding window method, the bottom-up method is adopted, and more sensitive time boundaries can be generated by learning convolutional networks. The model includes three steps:

(a) A series of videos is sparsely sampled to obtain a video frame and optical stream for each segment;
(b) This network generates action scores for video clips to judge whether there is action in the video. Here, the binary classifier of the TSN network is used to train two CNNs;
(c) Continuous high-score action fragments are grouped into action groups, and the threshold is set to allow the existence of some outliers. Multiple sets of thresholds are set to generate proposals of different granularity. This approach greatly reduces the number of proposals, covers more comprehensive actions, and simplifies the process of parameter tuning.

Lin et al.(Lin et al. 2018) suggests that the generation of candidate proposals requires proposals with precise time boundaries, and fewer proposals are needed to retrieve proposals. In addition, high-quality proposals need to be met that can cover real action areas with high recall rates and high time overlap, using fewer proposals to achieve higher recall and overlap rates. Therefore, Lin et al. proposed the boundary-sensitive network (BSN) in 2018, which can flexibly extract the action temporal candidate proposal. There are three main steps, as shown in Fig. 34:

Table 3 Summary of Anchor-based methods

Type	Year	Methods	Feature	Improve	Code
Anchor-based	2016	SCNN(Shou et al. 2016)	IDT	It creatively solves the problem of time action positioning in nonedited videos	https://github.com/zhengshou/scnn
	2016	PSDF(Yuan et al. 2016)	IDT	Proposed a Pyramid of Score Distribution Features that mitigates positional action and duration effects	-
	2017	CBR-TS(Gao et al. 2017b)	TS	Propose to use Cascaded Boundary Regression (CBR) to adjust temporal boundaries in a regression cascade, where regressed clips are fed back to the system for further boundary refinement	-
	2017	TURN(Gao et al. 2017a)	C3D	Proposed a novel architecture for temporal action proposal generation using temporal coordinate regression, proposed a new metric, AR-F, to evaluate the performance of TAP and compare AR-F with AR-AN and AR-N by quantitative analysis	-
	2017	R-C3D(Xu et al. 2017)	C3D	Propose an end-to-end model that can encode video streams using three-dimensional, fully convolutional networks	https://github.com/sunny xiaohu/R-C3D.pytorch
	2018	TAL-Net(Chao et al. 2018)	I3D	Using a multi-scale architecture that can accommodate extreme variation in action durations	-
	2021	AFSD(Lin et al. 2021)	I3D	Propose the first purely anchor-free temporal localization method	https://github.com/TencentYoutuResearch / ActionDetection -AFSD
	2021	SRF-Net(Ning et al. 2021)	C3D	A building block called Selective Receptive Fi-eld Convolution (SRFC) is dedicatedly designed and can adaptively adjust its receptive field size according to multiple scales of input information at each temporal location in the feature map	-
	2022	RCL(Wang et al. 2022a)	I3D	Proposed a continuous anchoring representation method, which unifies and extends existing anchor-based detectors into a continuous regression problem in 2D coordinates	-

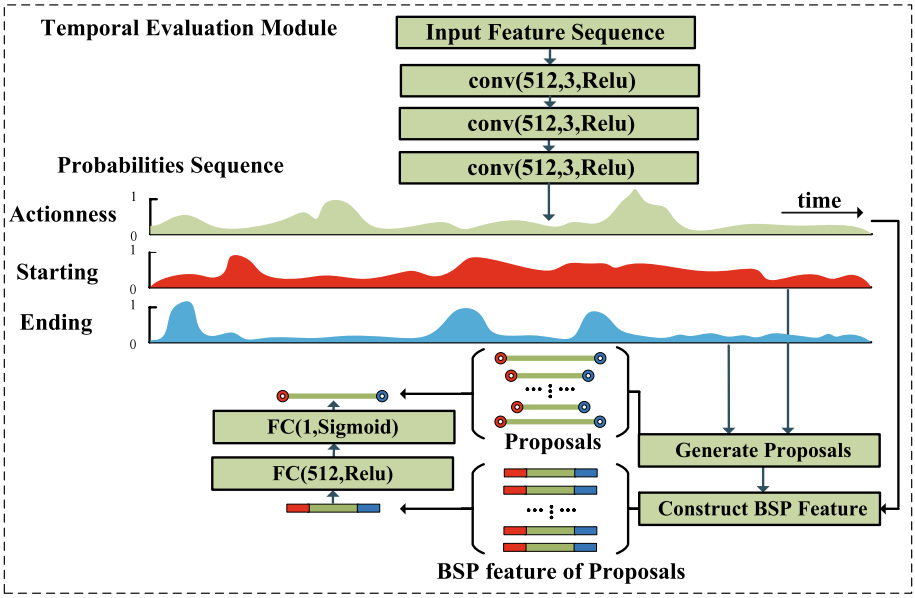


Fig. 34 Overview of BSN (Lin et al. 2018)

(a) Video feature extraction. The BSN uses a two-stream network to extract features, and each video unit is called a snippet;
(b) The boundary-sensitive network is used to generate action proposals and consists of three modules. The temporal evaluation module is used to generate the start probability sequence of an action, the end probability sequence of an action, and the probability sequence of an action. The proposal generation module extracts the action proposal by setting a threshold or a probability peak. The proposal evaluation module uses the MLP to score the confidence of each proposal;
(c) Non-maximum suppression is applied to overlapping parts to improve accuracy. The non-maximum suppression algorithm is Soft-NMS (Bodla et al. 2017), and the fractional attenuation function is used to suppress the redundant results.

This model is a multi-stage model, and is not the same network model. It lacks rich timing context information and is inefficient for constructing candidate proposal features separately from confidence evaluation. The proposed boundaries and timing are very flexible, and anchoring mechanisms are not suitable for bottom-up approaches such as BSN. In 2019, Lin et al. (Lin et al. 2019b) also proposed a boundary-matching network (BMN) to evaluate the confidence of dense distribution proposals. In the BM mechanism, a two-dimensional BM confidence graph represents the start and duration of all possible candidate proposals. The BM feature map contains abundant feature and context information, which optimizes the problem of insufficient semantic information in the temporal evaluation module in BSN. As shown in Fig. 35, an untrimmed video is input and is used to generate a boundary probability sequence using the boundary-matching network; the red line in the sequence predicts the start of the action, and the blue dashed line predicts the end. The boundary probability sequence can be matched

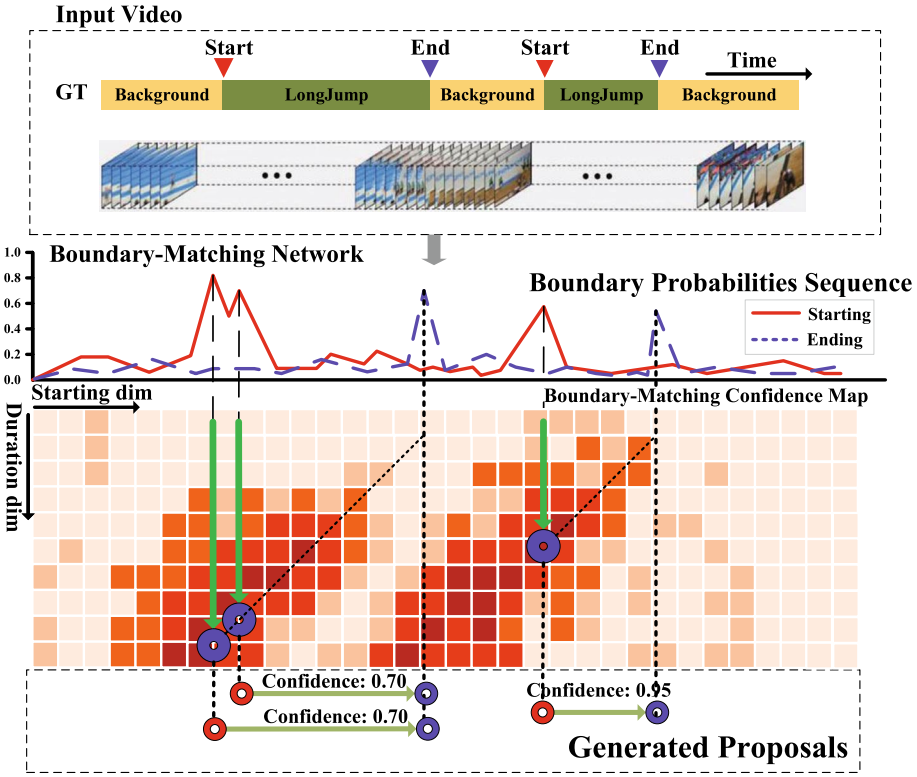


Fig. 35 Overview of BMN (Lin et al. 2019b)

with the boundary-matching confidence graph generated simultaneously, and the confidence of all proposals can be intensively evaluated.

Gong et al. (Gong et al. 2020b) have suggested that the time context is important. However, the durations of different action instances are different, and it is impossible to find a receptive field that fits every time. Similar to TAL-Net's multiscale receptive field, Gong et al. proposed TSA-Net in 2019, which uses a set of parameterized time convolution modules called multi-dilation temporal convolution blocks (MDCs) to handle all timescales. The model design is similar to (Lin et al. 2018; Chao et al. 2018) but is fundamentally different when simultaneously dealing with multiple timescales. To address the low accuracy of the boundary-based approach, TSA-Net detects three types of points simultaneously: the start point, the endpoint, and the midpoint of the action instance. The midpoint of the action instance implicitly encodes the candidate proposal, and a pair of start and end points is enabled only at the confident midpoint. TSA-Net outperformed competing methods on the large-scale benchmarks THUMOS14 and ActivityNet1.3 and recalibrated state-of-the-art performance.

While current proposal generation methods can generate precise action boundaries, little attention has been paid to the relationship between proposals. In 2021, Chen et al. (2021) proposed a unified framework for generating time boundary suggestions for graph convolutional networks based on boundary suggestion features, called boundary graph convolutional networks (BGCNs). BGCNs draw inspiration from boundary methods and use edge

graph convolution relays on the boundary proposals' feature. First, a BGCN uses a base layer to fuse the two-stream video features to obtain two branches of base features. The two branches of the basic features then enter the same structure of the proposed feature graph convolution network (PFGCN). The action PFGCN is used to extract the action classification score, and the boundary PFGCN is used to extract the end and start scores. In the proposed feature graph convolutional network, the proposed features are intensively sampled from the video features, and a proposed feature graph is constructed. Each proposal feature is taken as a node, and the relationship between proposal features is taken as an edge. Next, edge convolution is used for graph convolution, to map the relations into a 2D map score. As shown in Fig. 36, the method first uses a two-stream network to extract spatiotemporal features, after which features are sent to ActionPFGCN and BoundaryPFGCN, and then 2D map fractions are generated. Finally, the action classification score, the action regression score, the end score, and the beginning score are fused to produce a dense proposal. Soft NMS is then used to obtain the final real proposal, and then the proposed video clips are classified. Experimental results show that BGCN is an excellent proposal generator. In addition, BGCN has efficient action detection abilities, a model size of less than 2 MB, and a fast reasoning time.

As anchor methods lack flexible time boundaries, boundary methods have the problem of false positives in boundary prediction. Hsieh et al. (2022) proposed a contextual proposal network (CPN) centered on an RNN in 2022, encompassing two context-aware networks. The first mechanism, called feature enhancement, uses a similar inception module to capture multi-scale time contexts to produce robust representations of video footage. The second mechanism is the boundary scoring mechanism, which, for the first time, uses a bidirectional recursive neural network to capture the time context and formulate the exact boundary. This two-way time context helps to retrieve high-quality recommendations with low false-positives to override video action instances when generating and scoring proposals. Two challenging datasets, ActivityNet−1.3 and THOMAS-14, have demonstrated the effectiveness of CPN.

In 2021, Qing et al. (2021) proposed a temporal context aggregation network (TCANet) for high-quality proposal generation. Firstly, a local–global temporal encoder (LGTE) was proposed to capture both local and global time relations by channel grouping. The encoder consists of two sub-modules. Specifically, the input features are equally divided into N groups along the channel dimension after linear transformation. Then, a local temporal encoder (LTE) is designed to handle the first A groups for local temporal modeling. At the same time, the remaining N-A groups are captured by the global temporal encoder

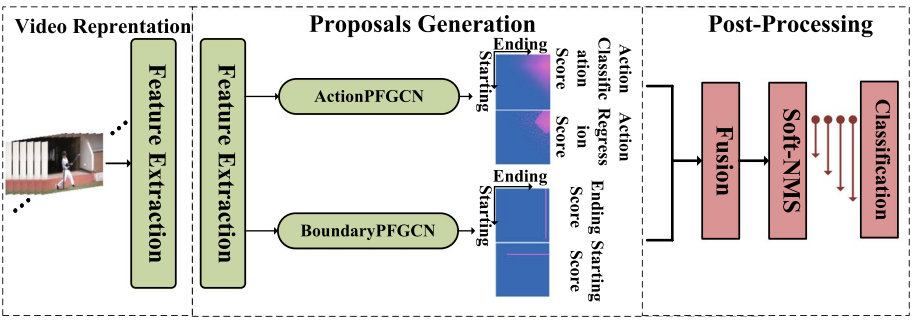


Fig. 36 Detail of the BGCN Workflow (Chen et al. 2021)

(GTE) for global information perception. In this way, LGTE is expected to integrate the long-term context of proposals using global groups while recovering more structure and detailed information from local groups. Second, the temporal boundary regressor (TBR) was proposed to exploit both the boundary context and internal context of proposals for frame-level and segment-level boundary regressions, respectively. Specifically, frame-level boundary regression aims to refine the start and end locations of candidate proposals with boundary sensitivity, while segment-level boundary regression aims to refine the center location and duration of proposals under the overall perception of proposals. Finally, high-quality proposals are obtained through complementary fusion and progressive boundary refinements. As shown in Fig. 37, given an untrimmed video, TCANet captures "local and global" time relationships in parallel through LGTE. In TBR, the inner and boundary contexts of the proposal are used for segment-level and frame-level boundary regression, respectively. Finally, the two regression outputs are fused to obtain the predicted results. Experimental results show that TCA-Net can significantly improve the performance of action proposals and action detection.

Liu and Wang (2020) proposed a progressive boundary refinement network (PBRNet) in 2020 to improve the accuracy and speed of TAD. Unlike most previous efforts, the entire network, including the feature extractor, is trained jointly. A three-step cascade regression pipeline is proposed to refine the boundary from coarse to fine for the detection process. Specifically, PBRNet consists of three main detection modules: coarse pyramid detection (CPD), fine pyramid detection (RPD), and fine-grained detection (FGD). CPD and RPD are anchoring-based detection systems in which two symmetrical feature pyramids are used to detect different action scales. FGD aims to refine the boundaries of action candidates by utilizing frame-level characteristics. In addition, three branches with different types of

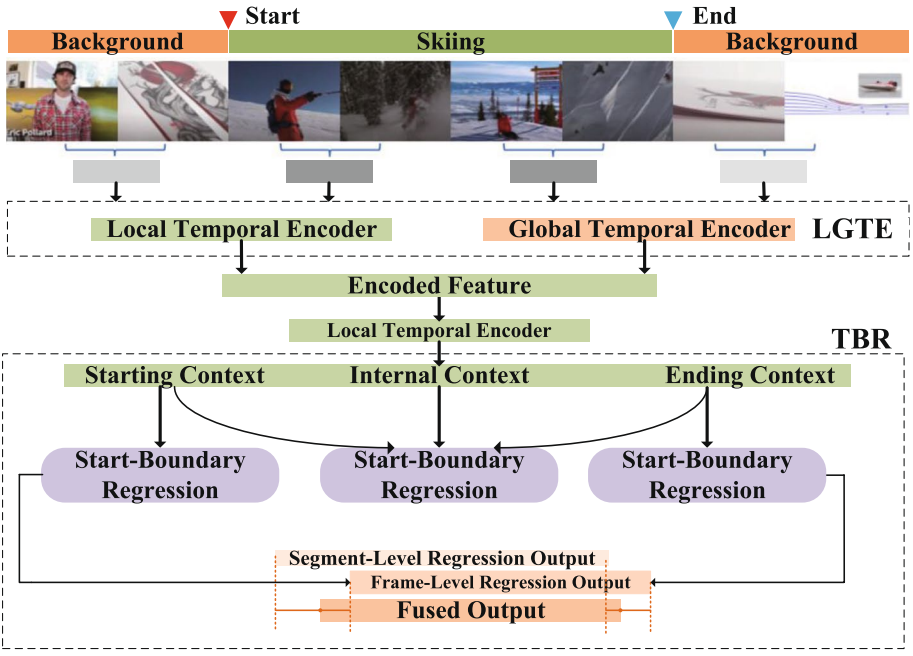


Fig. 37 Structure of the TCA network (Qing et al. 2021)

frame-level monitoring are used to enrich frame-level characteristics and update the classification score for each action instance. In particular, some learning strategies (such as progressive matching and preliminary anchor discarding) are used to cooperate with progressive learning. As a result, anchors are passed between adjacent modules for cascading regression and fusion of confidence scores from different modules for retrieval. The network has been proven to be effective using the ActivityNet-1.3 and THOMAS-14 datasets.

Boundary-based methods first generate the boundary probability sequence and then apply the boundary matching mechanism to generate the candidate proposal. TAG (Xiong et al. 2017) uses sparse sampling to generate action scores from video clips directly; BSN uses the boundary-sensitive network to determine the local time boundary with high probability and evaluate its global confidence. However, BSN ignores the global context of the video. To solve this problem, BMN uses the boundary matching network to aggregate the features of all proposals and simultaneously evaluate all proposals to capture the global context of the video. In spite of this, the methods of BSN and BMN do not attach importance to the global information of boundary prediction, which makes the location of actions with blurred boundaries inaccurate. DBG (Lin et al. 2020) solves this problem by using global information to predict the boundary probability. In addition, BGCN (Chen et al. 2021) and BCGCN (Bai et al. 2020) use graphs to model the relationship between the proposal's boundaries and content, where the proposal's boundaries and content are treated as nodes and edges, and their characteristics are updated via graph manipulation. CPN (Hsieh et al. 2022) and TCA-Net (Qing et al. 2021) use a context-aware network and a time context aggregation network to capture the time context. However, boundary-based methods usually ignore the scaling problem and use a fixed convolution-accepting field for all action instances. TSA-Net (Gong et al. 2020b) uses multi-extended time convolution to solve the timescale problem.

Boundary-based methods provide flexible boundaries, making up for the lack of flexible time boundaries for various action instance alignments in the anchored methods. Although the boundary context of the proposal considered in a boundary-based approach is sensitive to boundary changes, the resulting proposal is less reliable. Neither convolution nor global fusion can model time relations effectively. One-dimensional convolution operations (Lin et al. 2018, 2019b, 2020) lack flexibility in encoding long-term time relationships constrained by kernel size. The global fusion approach (Gao et al. 2020) ignores the various global dependencies of each time location and the implicit concern with local details, such as the local details of the boundary. The Table 4 is a summary of boundary-based methods, in which the feature types used by different network models and the corresponding improvements and code websites are given.

5.3 Query-based methods

The above two kinds of methods have been continuously improved and have proved effective with excellent performance. However, their limitations are difficult to eliminate. As a video becomes longer, the computing burden increases. In addition, it is susceptible to manual parameter settings, such as artificially designed anchor designs and confidence values. Recently, a new query-based approach has attracted the attention of researchers. It benefits from DETR (Carion et al. 2020) in the transformer, using a series of object queries instead of anchoring as a candidate object and creating a new view of action detection. This method uses only a small number of queries, so the network has a simple pipeline. The anchor frame and anchor point are removed based on fixed spatial position and the

Table 4 Summary of Boundary-based methods

Type	Year	Methods	Feature	Improve	Code
Boundary-based	2017	TAG Xiong et al. (2017)	TS	Can efficiently generate candidates with accurate temporal boundaries	-
	2017	SSN Zhao et al. (2017)	TS	Allows the framework to effectively distinguish positive proposals from background or incomplete ones	https://github.com/yjxiong/action-detection
	2018	BSN Lin et al. (2018)	TS	Use the boundary-sensitive proposal function to obtain more accurate boundaries and better retrieval quality	https://github.com/wzmsltw/BSN-boundary-sensitive-network
	2019	TSA-Net Gong et al. (2020b)	P3D	Convolution filters with different expansion rates are combined with expanding the receptive field effectively	-
	2019	BMN Lin et al. (2019b)	TS	Introduce the Boundary-Matching (BM) mechanism to evaluate confidence scores of densely distributed proposals	https://github.com/JJBOY/BMN-Boundary-Matching-Network
	2019	DBG Lin et al. (2020)	TS	Implements boundary classification and action completeness regression for densely distributed proposals	-
	2020	PBRNet Liu and Wang 2020)	TS	Three cascaded detection modules are used for fine action boundary location	https://github.com/canbaoburen/PBRNet
	2021	BGCN Chen et al. (2021)	TS	Uses graph convolution in boundary proposal features	-
	2021	TCA-Net Qing et al. (2021)	TS	Generate high-quality action proposals through "local and global" temporal context aggregation and complementary as well as progressive boundary refinement	
	2022	CPN Hsieh et al. (2022)	TS	Capture multi-scale time context using feature enhancement	-

method relies on a learnable vector for prediction instead. Because of their simplicity and flexibility, query-based methods have become a new solution for TAD.

Tan et al. (2021) argue that long-term time-context modeling is critical for proposal generation. Viewing the video as a time series and using the transformer architecture to model global one-dimensional dependencies can improve location performance. In their paper, Tan et al. propose two key problems, namely the feature full degree (Zhang and Tao 2012) and time boundary ambiguity (Satkin and Hebert 2010). In 2021, following the idea of the transformer, Tan et al. proposed RTD-Net, which is similar to the transformer architecture. There are three main improvements to DETR. To solve the problem of slow video processing, an attentive boundary module is used to replace the encoder in the transformer. In order to adapt to the problems of blurred time boundaries and sparse annotation and to reduce the single, strict evaluation criteria from the ground-truth, the relatively relaxed matching scheme is proposed. Finally, a three-branch detection head is designed for training and testing. As shown in Fig. 38, the entire model consists of three separate designs. The first is the boundary concern module for feature extraction, followed by the transformer decoder in the middle for querying and parallel decoding, and on the far right is a slack matcher for training label allocation. Many experiments have proven the effectiveness of RTD-Net. In addition, due to its simple design, RTD-Net is more efficient than the previous two design methods, eliminating the need for non-maximum suppression processing.

Although traditional methods have made great progress, problems such as multi-stage and manual design lead to the models' lack of efficiency and flexibility. In 2022, Liu et al. (2022a) proposed an end-to-end TAD model based on a transformer, called TadTR. TadTR adaptively extracts the time context required for each action prediction and updates the initial embedding with the extracted context. At its core is the time-variable attention module, which dynamically focuses on a sparse set of key clips in a video. In addition, a temporal deformable attention module is introduced into the model, which does not require NMS and can focus on a sparse set of keyframes adaptively. As shown in Fig. 39, using I3D encoded video features as input, after passing through a transformer encoder and decoder, features are input into the segment matching module for one-to-one ground-truth allocation. An action regression head is then used to evaluate the confidence of the predicted

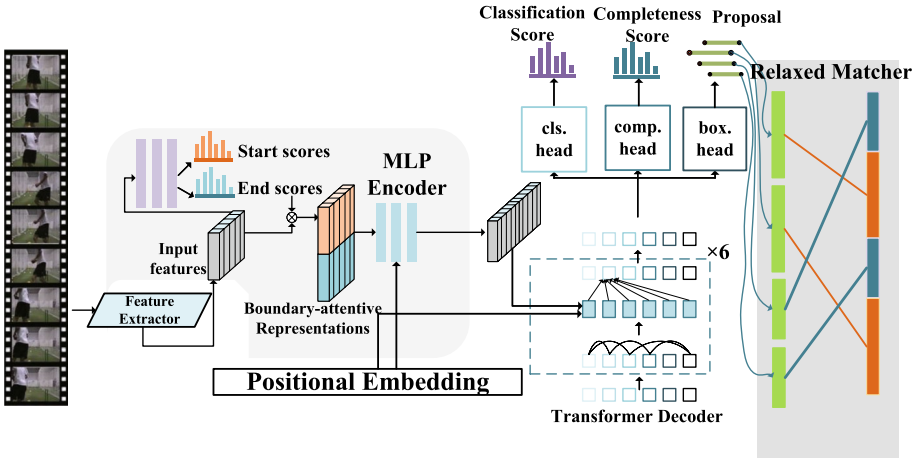


Fig. 38 Pipeline of RTD-Net (Tan et al. 2021)

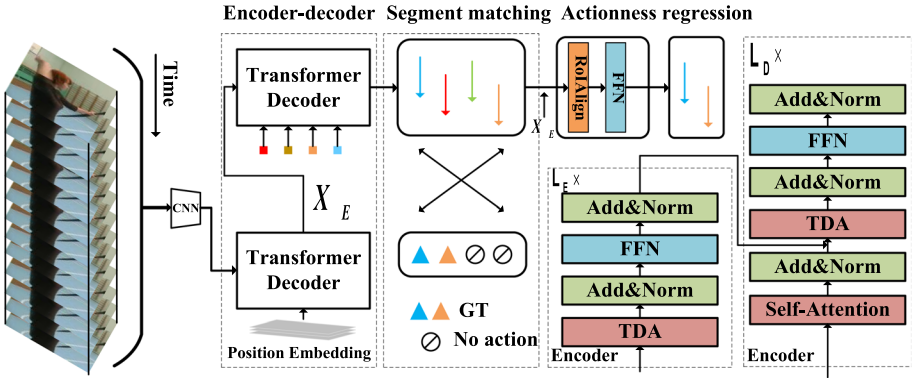


Fig. 39 The architecture of TadTR (Liu et al. 2022a)

fragment and finally output the predicted result of an action. TadTR can capture context information adaptively and is the first model used to study the adaptive context of TAD.

To address the problem of query-based methods being unable to build multi-scale feature maps, Wu et al. (2021) proposed an approach named SP-TAD in 2021, aiming at using sparse proposals and hierarchical features to exchange information so as to generate high-quality candidate proposals. SP-TAD is an end-to-end framework for feature extraction using I3D, which is composed of four main parts: 3D backbone network, time feature pyramid network, action detection header, and action classification header. SP-TAD proposes a simple, high-quality time sequence candidate generation framework, which removes the anchoring design and boundary design of the manual design by learning a small number of proposals. Sparse interaction is used to generate high-resolution features to improve model performance and put forward an iterative refinement strategy.

Methods such as DETR encounter problems when applied to TAD. Self-attention in the decoder does not fully explore interquery relationships because it is performed intensively across all queries. In addition, the DETR method may be affected by insufficient training in action classification. In order to alleviate these two problems, Shi et al. (2022) proposed two training loss functions in ReAct in 2022. The two training loss functions are called action classification enhancement (ACE) loss to promote classification learning. The first loss (ACE-enc) is applied to the feature input of the encoder. It aims to reduce the intra-class variance and inter-class similarity of action instances. This loss improves the discriminability of video features related to the performance category, thus benefiting the classification. Meanwhile, the second loss (ACE-dec) is proposed as a classification loss in the decoder, which considers both predictions and ground truth segments for action classification. It adds training samples and generates stable learning signals for the classifier.

The emergence of query-based methods introduced the encoder-decoder framework of the transformer into temporal action detection. The detected action fragments were modeled as a fixed number of learnable query vectors. In contrast to the above two methods, their performance depends largely on elaborate anchor placement or complex boundary-matching mechanisms that are developed using prior human knowledge and require specific adjustments. In contrast, query-based approaches use only a small set of queries, have simple pipelines, and are free of manual design. The Table 5 is a summary of query-based methods, which presents the feature types used by different network models and the corresponding improvements and code websites.

Table 5 Summary of Query-based methods

Type	Year	Methods	Feature	Improve	Code
Query-based	2021	SP-TAD Wu et al. (2021)	I3D	Sparse proposals are introduced to interact with hierarchical features	https://github.com/wjn922/SP-TAD
	2021	RTD-Net Tan et al. (2021)	I3D	Improve the original Detr with the help of transformer	https://github.com/MCG-NJU/RTD-Action
	2022	TadTR Liu et al. (2022a)	I3D	Mapping a group of learnable embedded parallel to action instances based on transformer	https://github.com/xlliu7/TadTR
	2022	ReAct Shi et al. (2022)	I3D	Propose a Segment Quality to predict the localization quality of each action query	https://github.com/ssssste/React

6 TAD according to the learning method

Data annotation in video action localization is mainly performed manually by humans, which leads to some difficulties. Due to the high time cost, the difficulty in assigning precise boundaries to each action, and the inevitable human subjectivity, in order to reduce the cost, time, and manpower, some researchers have proposed weak supervision and unsupervised learning. This section introduces learning methods, summarizes the main fully supervised learning methods at present, and focuses on weakly supervised learning methods. Below are descriptions of the three types of learning methods (Fig. 40).

6.1 Full supervision

Fully supervised learning refers to obtaining the optimal model through the existing training samples and then mapping the input to the corresponding output label for judgment. Research includes the prediction problem, where the input and output variables are both continuous, and the classification problem of the output finite discrete variables. In video understanding, full supervision uses video-level category tags and time-stamped information tags of action clips. There are many methods, which can be divided into the three methods described above in terms of design ideas: anchor-based (another derived branch can be called anchor-free), boundary-based, and query-based. The idea of an anchor core-based method is derived from target detection. Early anchor core-based methods mainly use a fixed sliding window to generate candidates. The most typical sliding window method is the SCNN model proposed by Shou et al. (2016), which adopts multiple frame-rate windows to intercept video clips and uses a three-stage SCNN for processing. However, due to excessive overlap, the SCNN model has some problems, such as redundancy and poor

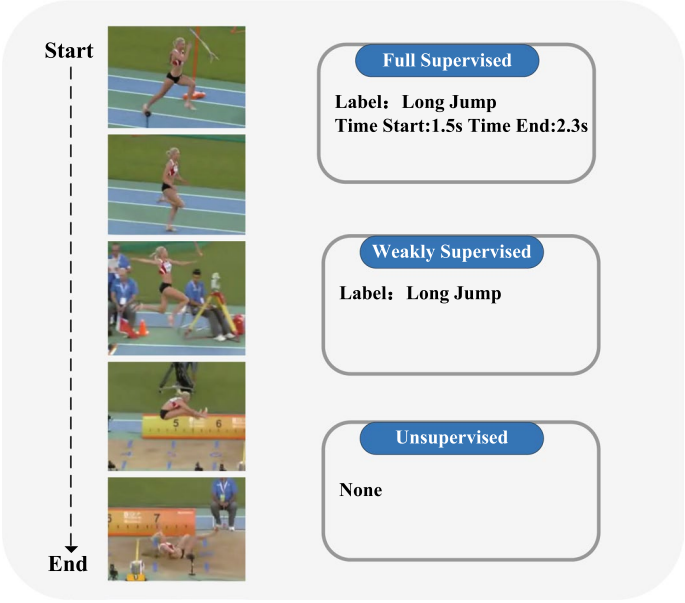


Fig. 40 Three types of supervised learning and notes (Baraka and Mohd Noor 2022)

calculation accuracy. In order to solve the accuracy problem, Gao et al. (2017a) proposed the TURN model, which uses the idea of boundary regression in Faster-RCNN (Ren et al. 2015) for reference. By introducing the coordinate regression offset, the boundary error caused by the fixed candidate box can be modified to a certain extent, but the problem of boundary imprecision is not solved. In addition, Long et al. (2019) proposed the GTAN model in 2019, which uses a Gaussian kernel to simulate the time window and optimize each generated action proposal. Adding Gaussian kernels can represent action proposals of different sizes, and the corresponding Gaussian curves can show the context relationships generated into action proposals.

Another alternative to the generation of an anchor-based candidate proposal is the generation of a boundary candidate proposal based on action probability distribution by dichotomizing the action and background of the video clip or single frame, and obtaining the probability curve of the video area as the action. Xiong et al. (2017) proposed the TAG algorithm in 2017. They designed a learning-based bottom-up scheme to generate candidate proposals that are more sensitive to time boundaries than the sliding windows generated by conventional schemes. Zhao et al. (2017) proposed the SSN algorithm with the same design idea as Xiong et al., and both of them adopted the classic "proposal + classification" paradigm. A structured time pyramid is introduced to generate a global representation of the entire scheme, and a decomposed discriminant model is added to combine the classification of action categories and determine the completeness of the proposal. These proposals work together to output only the complete action instance. In order to improve the quality of the generation proposal, Lin et al. proposed BSN (Lin et al. 2018) in 2018, which adopts a local-to-global approach, locally combines high-probability boundaries into proposals, and uses proposal-level features to retrieve candidate proposals globally. However, the steps required to generate the proposal were tedious, the network model was complex, and the constructed proposal characteristics were too simple to capture the context information. Therefore, an improved BMN (Lin et al. 2019b) model based on BSN was proposed in 2019. BMN uses a boundary-matching (BM) mechanism to evaluate the reliability of the dense distribution proposal. Similarly, Lin et al. (2020) proposed a dense boundary generator (DBG) to use global suggestion features to predict boundary graphs and explore action perception features for action integrity analysis.

In recent years, new query-based approaches have emerged, which are worthy of attention as the number of such methods is currently very small. The most representative method is the RTD-Net proposed by Tan et al. (2021) in 2021, which introduces three important improvements to DETR; it solves the essential visual difference between time and space. Liu et al. (2022a) proposed TadTR in 2022. TadTR is an end-to-end framework based on the transformer, which maps a set of learned embeddings to parallel action examples. Although the TAD algorithm is still in continuous development, its accuracy rate has reached 57.1% with an IoU of 0.5; compared with image processing, there is still a long way to go. The Table 6 is a summary of fully supervised method in the THUMOS14 dataset.

6.2 Weak supervision

The TAD of the full supervision method requires action labels and time boundary annotations for all actions. Because of the high cost of data marking in TAD and people's subjectivity when identifying the exact start time and end time of the action, errors can easily occur in the annotation. Therefore, fully supervised learning is labor-intensive and

Table 6 Performance of each fully supervised method in the THUMOS14 dataset

Type	Year	Method	Feature	Map@IoU(%)				
				0.3	0.4	0.5	0.6	0.7
Anchor-based	2016	SCNN Shou et al. (2016)	IDT	36.3	28.7	19.0	–	–
	2016	PSDFYuan et al. (2016)	IDT	33.6	26.1	18.8	–	–
	2017	CBR-TS Gao et al. (2017b)	TS	50.1	41.3	31.0	19.1	9.9
	2017	TURN Gao et al. (2017a)	C3D	44.1	34.9	25.6	–	–
	2017	R-C3D Xu et al. (2017)	C3D	44.8	35.6	28.9	–	–
	2018	TAL-Net Chao et al. (2018)	I3D	53.2	48.5	42.8	33.8	20.8
	2021	AFSD Lin et al. 2021)	I3D	67.3	62.4	55.5	43.7	31.1
	2021	SRF-Net Ning et al. (2021)	C3D	56.5	50.7	44.8	33.0	20.9
	2022	RCL Wang et al. (2022a)	I3D	70.1	62.3	52.9	42.7	30.7
Boundary-based	2017	TAG Xiong et al. (2017)	TS	48.7	39.8	28.2	–	–
	2017	SSN Zhao et al. (2017)	TS	51.0	41.0	29.8	–	–
	2018	BSN Lin et al. (2018)	TS	53.5	45.0	36.9	28.4	20.0
	2019	TSA-Net Gong et al. (2020b)	P3D	61.2	55.9	46.9	36.1	25.2
	2019	BMN Lin et al. (2019b)	TS	56.0	47.4	38.8	29.7	20.5
	2019	DBG Lin et al. (2020)	TS	57.8	49.4	42.8	33.8	21.7
	2020	PBRNet Liu and Wang (2020)	TS	58.5	54.6	51.3	41.4	29.5
	2021	BGCN Chen et al. (2021)	TS	60.8	53.3	44.8	34.1	23.3
	2021	TCA-Net Qing et al. (2021)	TS	60.6	53.2	44.6	36.8	26.7
	2022	CPN Hsieh et al. (2022)	TS	68.2	62.1	54.1	41.5	28.0
Query-based	2021	SP-TAD Wu et al. (2021)	I3D	69.2	63.3	55.9	45.7	33.4
	2021	RTD-Net Tan et al. (2021)	I3D	58.5	53.1	45.1	36.4	25.0
	2022	TadTR Liu et al. (2022a)	I3D	62.4	57.4	49.2	37.8	26.3
	2022	ReAct Shi et al. (2022)	I3D	69.2	65.0	57.1	47.8	36.5

subjective. In order to reduce the model's need for detailed annotation of datasets, researchers began to explore weakly supervised learning methods. Weakly supervised learning can effectively deal with the problems of missing or inaccurate labels in machine learning and can match the TAD task.

In the process of training, the weak supervision model only requires the video-level label and autonomously learns the action boundary in the video. The most common way to do this is to use the attention mechanism to focus on distinguishing segments and to convert prominent episode-level features into video-level features. Discriminative fragments are obtained through the two methods discussed in this section: multi-instance learning and direct localization. In addition, two common challenges in weakly supervised learning are discussed: complete action modeling and action-context separation.

6.2.1 Multi-instance learning

Multi-instance learning (MIL) is a form of weakly supervised learning. The training instances are arranged in groups, called bags, and provide labels for the entire bag. Supervision is provided only for the complete set of products, and no separate labels are provided for the instances contained in the bag. This formulation of the problem has attracted a lot

of attention in the research community, especially in recent years, when the amount of data needed to solve big problems has increased exponentially. Large amounts of data require increasing amounts of tagging work.

In TAD, each complete video is treated as a bag of action instances. A single confidence score for each video is obtained by calculating the loss per bag. The confidence score for each category is calculated as the average of the first k activation scores for that category in the time dimension. The video-level confidence score for class c is defined as s^c. The probability distribution p^c is calculated by applying the softmax function to the s^c fraction in the class dimension. MIL loss is the cross-entropy loss applied to all videos and all action classes. For video is i and action type is c, p_i^c is the class probability fraction and X_i^c is the binary label of the normalized ground-truth. The numbers of action categories and videos are represented by n_c and n. Formula 2 gives the loss function for multi-instance learning.

$$L_{MIL} = \frac{1}{n}\sum_{i=1}^{n}\sum_{C=1}^{n_c} -y_i^C log(p_i^C), p^C = \frac{exp(s^C)}{\sum_{C=1}^{n_C} exp(s^C)} \tag{3}$$

The UntrimmedNet proposed by Wang et al. (2017) in 2017 is a weak supervision network that uses a multi-instance learning framework for the first time and does not depend on the selection of a video feature extraction network. It is also the first time an action recognition model can be learned from untrimmed video without needing time annotations for action instances. UntrimmedNet coupling classification and selection modules, implemented through feedforward networks, comprise an end-to-end model with even better performance than the fully supervised model. The classification module uses linear mapping to a multidimensional score vector, which is then passed using softmax. The selection module uses two selection mechanisms: hard selection based on multi-example learning and soft selection based on attention modeling. In hard selection, inspired by multi-instance learning, a subset of k clip proposals (instances) is identified for each action class, then the first k instances with the highest classification scores are selected, and finally the average is taken among these selected instances. In soft selection, for candidate proposals that do not need to be action-related, attention can be used to highlight different candidate proposals to distinguish (suppress) candidates that are background proposals. As shown in Fig. 41, first, sampling is performed from unclipped videos, and then the clipping scheme is fed into the

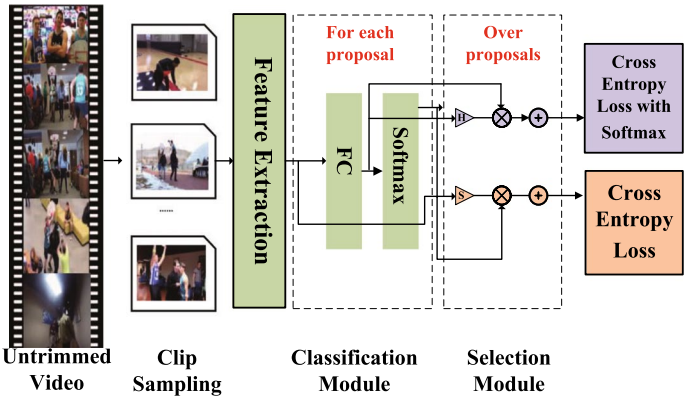


Fig. 41 Pipeline of learning from untrimmed videos (Wang et al. 2017)

pre-trained network for feature extraction. Secondly, UntrimmedNet uses a classification module to independently perform action recognition for each clip suggestion and proposes a selection module to detect or sort important clip suggestions. Finally, a video-level prediction is obtained by combining the outputs of the classification module and the selection module.

Lee et al. (2020b) proposed a background suppression network based on BaSNet in 2019 to solve the problem of background classification in multi-instance learning, although the background does not belong to any class. However, when the background is considered a category of action, the loss can be reduced, even though the background does not have action characteristics. Further, inconsistencies between action and background can lead to errors and performance degradation, and it is clear that multi-instance learning cannot achieve this process. Through the background suppression network, the background is grouped into an auxiliary category and contains an asymmetric two-branch weight-sharing architecture with filtering modules and comparison targets. The two branches are the basic branch and the inhibitory branch (Narkhede et al. 2022). The suppression branch begins with the filter module, which is expected to attenuate the input features from the background frame. Unlike the basic branch, the goal of the suppression branch is to minimize the background category score of all videos while optimizing the original goal of the action category. By sharing their weights, the two branches take a feature map and generate CAS to predict video-level scores. Because the two branches share weights, they cannot optimize the two comparison targets at the same time with the same input. To address this limitation, the filter module learns how to suppress activation from the background. As shown in Fig. 42, (a) is the feature extractor, which inputs RGB and optical flow features into two branches, namely (b) the basic branch and (c) the suppressed branch. The two branches share weights and generate a frame class activation sequence (CAS) to classify the video as positive samples of the action and background classes. Experimental results show that BasNet is effective for background suppression, and its performance is improved in comparison to other methods.

UntrimmedNet (Wang et al. 2017) suggests learning a selection module for detecting important fragments; BasNet (Lee et al. 2020b) uses a background suppression network that regards the background as an action category; W-TALC (Paul et al. 2018) introduces

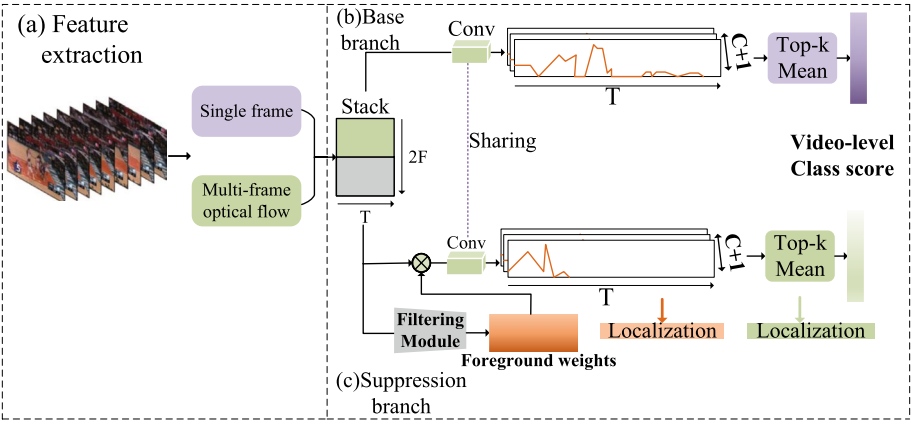


Fig. 42 Overview of the BaSNet (Lee et al. 2020b)

co-active similarity loss and optimizes weakly supervised time action detection together with cross-entropy loss. However, the relationship between video interference and movement has not yet attracted attention. STAR, proposed by Xu et al. (2019) in 2018, is another end-to-end framework inspired by multi-instance tags. The framework is a generative representation of actions called instance patterns aggregated by attentional mechanisms, and learns temporal relationships between them through a recursive neural network (Shi et al. 2015). Actions can be predicted and localized in time by designing a scoring term fused with attention weights (ST-GradCAM).

In 2021, Huang et al. (2021) proposed an FAC-Net framework based on the I3D trunk. Three branches are attached to the framework: the class foreground classification branch, the class-independent attention branch, and the multi-instance learning branch. As shown in Fig. 43, the video is processed as RGB and the optical stream is used as input. The first branch, CA, was designed to model the relationship between action and foreground. Its effect is similar to that of noise contrast estimation (NCE) (Oord et al. 2018; Gutmann and Hyvärinen 2010), which maximizes the lower limit of mutual information (MI) between foreground features and real ground action features, thus achieving better foreground-background separation. The second branch, CW, introduces an independent attention mechanism that models the reverse foreground-to-action relationship to complement the first branch and establish foreground action consistency. In addition, it is able to learn semantically meaningful foreground features. The third branch is an MIL-like pipeline for further improving video classification and facilitating classroom attentional learning in the CW branch first. In addition, the class-independent attention branch and the multi-instance learning branch are used to regularize the consistency of the front action, which helps the model to learn meaningful prospect classifiers. In each branch, a hybrid attention mechanism is introduced. The mechanism calculates multiple attention scores for each segment to focus on distinguishing and less distinguishing segments, thus capturing the full action boundary.

General MIL learning is based on two characteristic modes, namely RGB frames and optical flow, which are fused in two ways. The early fusion approach connects RGB and optical flow features before feeding them into the network, while the late fusion approach calculates the weighted sum of their respective outputs before generating recommendations for action. Although these developments have had some success, there are two unavoidable challenges to weakly supervised learning. First of all, it is difficult to exclude the possibility of a false-positive action proposal. Since there is no frame-level label, it is easy to

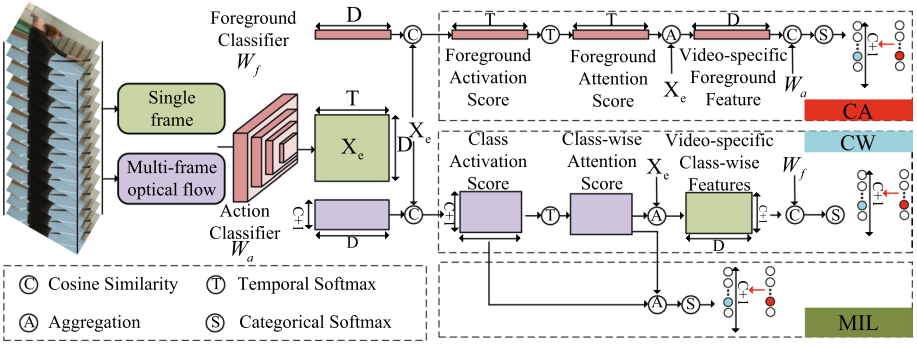


Fig. 43 Overview of the FAC-Net (Huang et al. 2021)

misjudge non-action instance content in the video as an action instance only by the video-level label corresponding to the action instance. Therefore, it is necessary to emphasize the supervision of fineness. Another problem is in the formulation of action proposals, which are generated by thresholding the activation sequence with a fixed threshold that is empirically preset. This has a significant impact on the quality of recommendations for action: a high threshold may lead to incomplete recommendations for action, while a low threshold may lead to false positives. To solve the above two problems, Zhai et al. (2020) proposed a two-stream consensus network (TSCN) in 2020. The authors designed an iterative refinement training scheme, where a frame-level pseudo-ground truth is generated from the late fused attention sequence and used as more accurate frame-level supervision to iteratively update both streaming models. In order to improve the quality of the generated proposal, an attention-normalized loss function is used. This improves the quality of the proposals generated by the threshold method by forcing the attention mechanism to make only limited choices as in binary methods.

6.2.2 Direct localization

Some methods generate action proposals by directly processing attention scores. As mentioned above, UntrimmedNet (Wang et al. 2017) sets a threshold to locate actions. Thresholds handle time segments independently and are not robust to noise in the class activation diagram. Instead of changing the algorithm (Song et al. 2014) and relying on external data (Singh et al. 2016), the hide-and-seek method proposed by Kumar Singh and Jae Lee (2017) in 2017 to directly changes the input image. By randomly hiding the image and forcing the network to find other relevant parts, it can be easily moved to different neural networks and tasks. In the TAD task, the hidden picture part is changed to hide the random frame sequence, thus forcing the network to learn the frames associated with the action. This is not appropriate for TAD tasks. Figure 44 shows an example of the hide-and-seek algorithm randomly hiding an image.

In 2018, Shou et al. (2018) proposed AutoLoc and conducted direct boundary prediction by predicting each action instance's central position and duration and obtaining the outer boundary by inflating the inner boundary. To solve the problem of training the boundary prediction model without real ground boundary annotation, they designed a

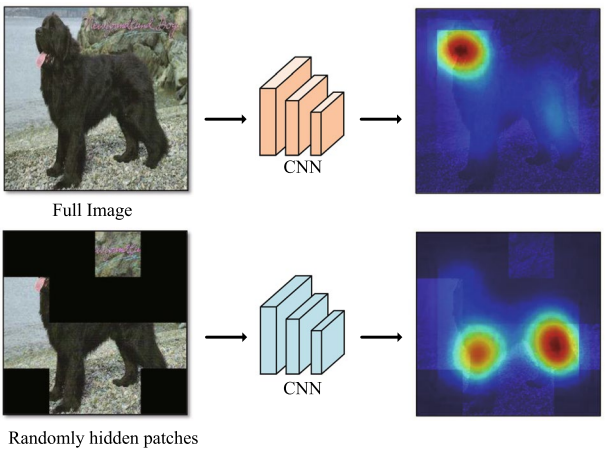


Fig. 44 Main idea of hide and seek (Kumar Singh and Jae Lee 2017)

novel outer-inner-contrastive (OIC) loss. OIC loss encourages high activation in the inner region. It punishes high activation in the outer region, allowing for the ideal positioning of significant intervals on the CAS, which can be well aligned with the ground truth segment.

In 2018, Nguyen et al. (2018) proposed a sparse temporal pooling network (STPN) that measures the video-level classification error and sparsity of selected segments by using the loss function to learn useful video clips from the sparse subset of gesture recognition for each video. STPN uses the time class activation graph to produce a one-dimensional time proposal of the target action. As shown in Fig. 45, the pre-training network is first used to extract the feature representation for a group of uniformly sampled video clips. The attention module is then used to calculate the class-independent attention weight for each segment, which is used to generate a video-level representation through a weighted time average pool. Finally, the representation of the classification module is given, which can be trained using the planned cross-entropy loss with a video-level label.

The above WS-TAL method (Wang et al. 2017; Kumar Singh and Jae Lee 2017; Nguyen et al. 2018) locates actions by direct thresholding of the classification scores of each segment. Therefore, these fragments are treated independently, and their temporal relationships are ignored. In fact, true action boundaries often depend heavily on temporal contrasts between these segments, such as temporal discontinuities and sudden changes. CleanNet, proposed by Liu et al. (2019c) in 2019, calculates one action score and two boundary scores for each action proposal, respectively representing the likelihood that an action proposal contains a particular action and the agreement that an action proposal begins and ends at the edge of a particular action. By combining action, opening, and closing scores, the new proposed action evaluator provides a comprehensive "comparison score" that measures both the content and the completeness of the proposed action. The

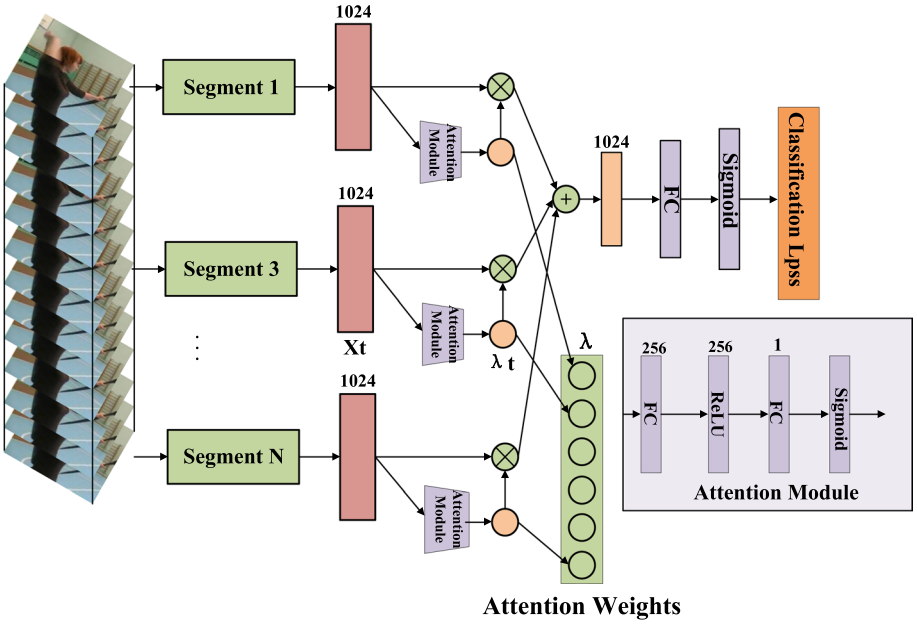


Fig. 45 Overview of the STPN (Nguyen et al. 2018)

framework is trained by maximizing the average contrast score of proposals and penalizing scattered short proposals, thereby improving the integrity and continuity of proposals.

6.2.3 Complete action modelling

A complete action sequence contains the same sub-action time series. For example, the long jump consists of three sub-sequences: run-up, take-off, and landing. In full supervision, it is easy to learn the complete action from the comments. However, in weak supervision, due to the lack of time boundary annotation, the integrity modeling of the action needs additional research.

The CMCS proposed by Liu et al. (2019a) in 2019 uses a multi-branch network and diversity loss to ensure dissimilarity between the class activation sequences output by different branches, thus training each branch to locate different parts of the action. Aggregating activations can retrieve the complete action from multiple branches. Class activations are then concentrated over time, resulting in a video-level distribution of categories.

HAM-Net was proposed by Islam et al. (2021) in 2021, wherein the mixed attention mechanism was adopted to solve the problem that when an action contains multiple sub-actions, the multi-instance learning framework can only detect specific sub-actions. As shown in Fig. 46, the entire framework contains a classification branch for predicting class activation scores for action instances (including background instances). A branch of attention is used to predict the score of a video clip. The mixed attention mechanism feeds RGB and optical flow features to the classification and attention branches, respectively. It is adjusted by three episode-level attention scores, namely the semi-soft attention score, soft attention score, and hard attention score, and is temporarily aggregated to generate video-level class scores. HAM-Net trains the network with four types of attention-directed losses: base classification loss (BCL), soft attention loss (SAL), semi-soft attention loss (SSAL), and hard attention loss (HAL). In an innovative move, using hard attention to capture complete examples of action reduces the more recognizable parts of the video and focuses on the less recognizable parts. This is done by calculating video clips' semi-soft and hard attention scores.

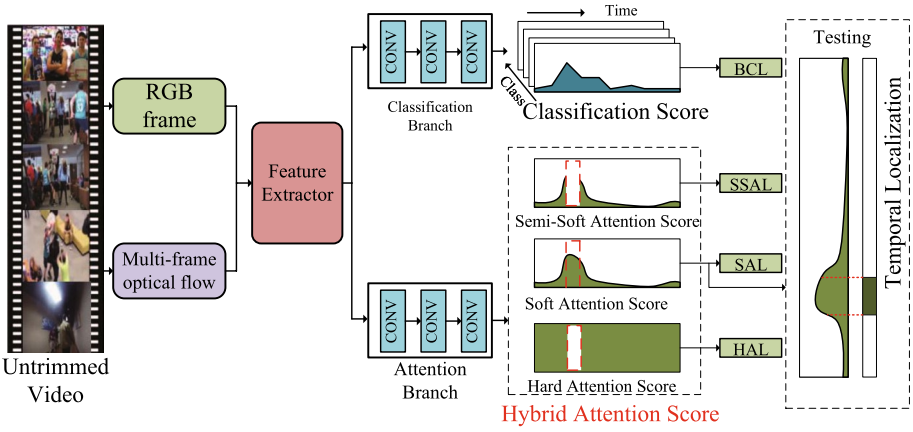


Fig. 46 Overview of the HAM-Net (Islam et al. 2021)

Huang et al. (2020) proposed a clustering loss function based on the co-occurrence GCN of RPN proposed in 2020. Clustering loss can push the features of an action to its corresponding prototype, thus generating clustering features, which can help detect complete action instances. Since different action instances may exhibit different speeds of action, especially slow action, slow action is defined as an action that is slower than normal. This is ignored in many papers, in which a common channel is used to extract features from video frames sampled at a fixed rate. As a result, the WTAD framework following this pipeline is difficult to locate slow action. Therefore, Sun et al. (2022a) proposed the slow-motion enhanced network (SMEN) in 2022, which exploits the related functions of slow motion through two modules. A mining module for generating masks is used to filter out the masks of slow-motion-related features from the entire video feature. A positioning module is used to predict the time boundary of an action using overall video features and intentional slow-motion features. As shown in Fig. 47, the mining module performs sampling operations on the original features and uses the CAS generation backbone to generate CASsub, which is then used to generate the mask to generate the enhanced slow-motion feature Xslow. The positioning module consists of two branches, as shown in the figure, one of which (i.e., the branch centered on normal motion) takes the raw video feature as input, while the other (i.e., the slow-motion branch) takes the enhanced slow-motion feature Xslow as input. The CAS and attention weights generated by these two branches are combined via the fusion module to output the final prediction.

Existing WSTAL methods usually adopt early fusion or late fusion. This simple cascading or fusion method is indirect and is not allowed, leading to many incorrect detection results and a failure to make full use of the action information. Therefore, Cao et al. (2022) proposed a deep motion prior network (DMP-Net) in 2022 to make full use of optical flow modes by learning an effective context-dependent action representation. This movement indicates global awareness, focusing on movements of interest regardless of background and irrelevant movements. XE losses are designed to measure the performance of classification models and are inherently incompatible with positioning tasks. Therefore, a plug-and-play loss function is proposed to replace the traditional XE loss function under weak supervision.

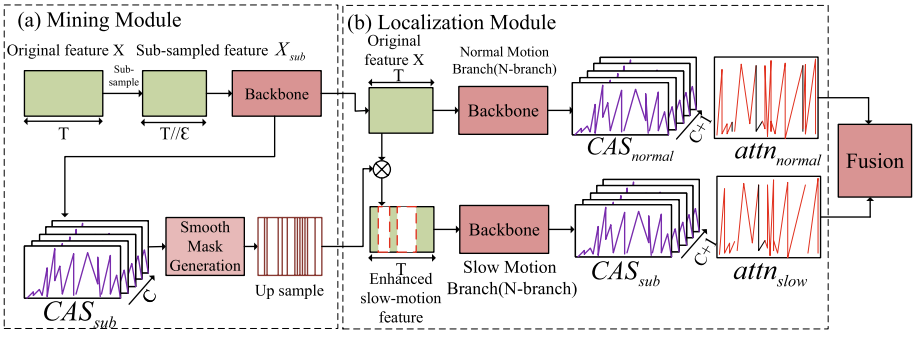
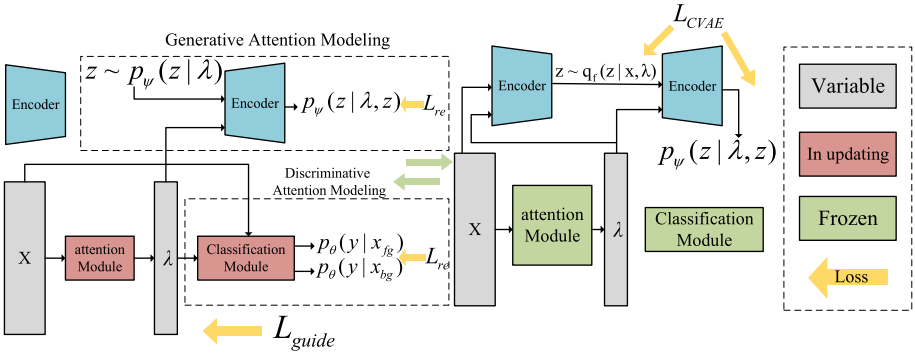


Fig. 47 Overview of the SMEN (Sun et al. 2022a)

6.2.4 Action-context separation

Frames adjacent to the beginning and end of an action cannot be considered part of the action. Action-context confusion can arise due to the absence of frame-level labels; context frames near an action fragment are often identified as action frames themselves because they are closely related to a particular class. To solve this problem, Shi et al. (2020) proposed an attention-generating mechanism in 2020 to focus on framing representations with framing attention as the core. In addition, by establishing a graphical model, it was proved that the video action location problem is related to the traditional classification and representation model. The proposed model consists of two parts: discriminative and generative attention modeling (DGAM). On one hand, discriminative attention modeling trains a classification model on temporally pooled features weighted by the frame attention. On the other hand, a generative model, i.e., a variational conditional auto-encoder (VAE), is learned to model the class-agnostic frame-wise distribution of representation conditioned on attention values. By maximizing the likelihood of the representation, the frame-wise attention is optimized accordingly, leading to good separation of action and context frames. As shown in Fig. 48, the model is divided into two alternating stages, (a) and (b), for training. In phase (a), generative models (CVAE) are frozen. The attention module and classification module are updated with classification-based discriminant loss Ld, representation-based reconstruction loss Lre, and regularized loss Lguide. In phase (b), the attention and classification modules are frozen. CVAE is trained to use losses to reconstruct representations of frames with different λ values.

Nguyen et al. (2018) proposed a new weakly supervised learning method in 2018. The model selects a sparse subset of useful video clips so that loss functions can be used for action recognition. The loss function measures video-level classification errors and the sparsity of the selected footage. The discriminant frame describing the action instance is highlighted using the attention weight, and the background frame is deleted. Due to the lack of a frame-level framework, the weakly supervised TAD approach is not effective in clearly separating action from context. Zhai et al. (2022) improved on the basis of TSCN and proposed an adaptive two-stream consensus network (A-TSCN) in 2022. The A-TSCN proposes an adaptive normalized loss of attention to better distinguish between action and background. Adaptive attention normalization loss automatically distinguishes between



(a) Updating attention/classification modules (b) Updating CVAE

Fig. 48 Overview of the Shi et al (2020)

action clips and background clips based on video attention distribution. By maximizing the difference between the attention values of the action fragment and the background fragment, the adaptive attention normalization loss facilitates the precise localization of the action boundary.

However, the current TAD approach based on weak supervision does not take full advantage of the short-term consistency between successive frames and the long-term continuity within the action, resulting in reduced accuracy in detecting action boundaries in the untrimmed video. Li et al. (2022a) introduced a superframe-based temporal proposal (SFTP) in 2022. Superframes are used to replace successive frames that change slowly at the same stage because they are consistent and their features are redundant. Having a superframe as the basic unit of video rather than a single frame avoids the need to categorize consistent frames into different categories. Moreover, because the salient parts of the segment are easier to identify, the recognition results tend to be trapped in the recognizable action segment rather than the entire instance. Li et al. also devised a scaling normalization strategy that isolates the effects of different scaling proposals while detecting various actions in a video. As shown in Fig. 49, the obtained superframe is input into the SFTP module; the green line is the classification branch in the prediction network, while the blue line is the detection branch. Finally, the SFTP scores of the two branches are obtained through global normalization for prediction.

The Table 7 is a summary of weakly supervised method in the THUMOS14 dataset.

6.3 Unsupervised

Unsupervised pre-training has attracted considerable attention in recent years for its potential to mine large amounts of unlabeled data. Contrast learning (Oord et al. 2018; Chen et al. 2020a, b; Grill et al. 2020; He et al. 2020) focuses on one of the most popular directions in

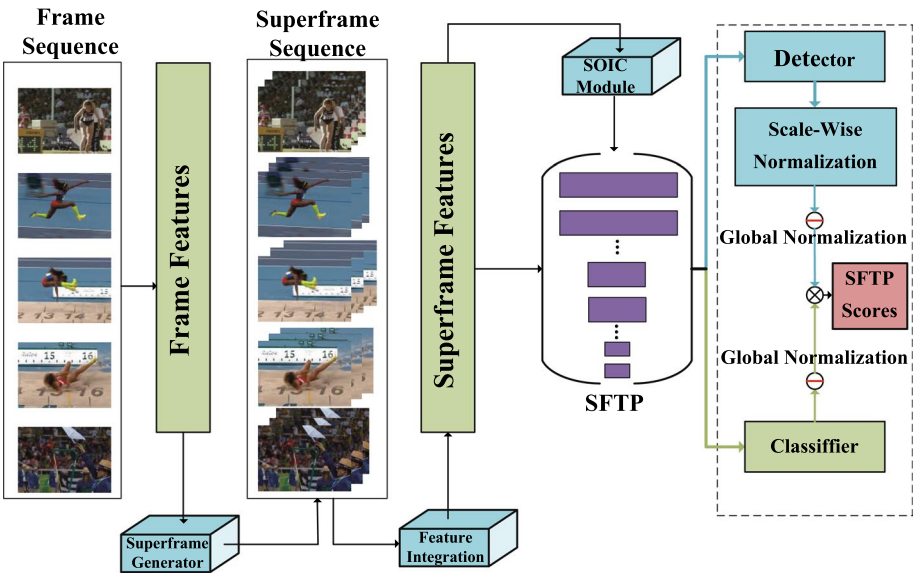


Fig. 49 Overview of the SFTP (Li et al. 2022a)

Table 7 Performance of each weakly supervised method in the THUMOS14 dataset

Type	Year	Method	Feature	Map@IoU(%)				
				0.3	0.4	0.5	0.6	0.7
Weekly	2017	Hide and SeekKumar Singh and Jae Lee (2017)	–	19.5	12.7	6.8	–	–
	2017	UntrimmedNetWang et al. (2017)	–	28.2	21.1	13.7	–	–
	2018	W-TALC Paul et al. (2018)	I3D	40.1	31.1	22.8	–	7.6
	2018	AutoLoc Shou et al. (2018)	UNT	35.8	29.0	21.2	13.4	5.8
	2018	STPN Nguyen et al. (2018)	I3D	35.5	25.8	16.9	9.9	4.3
	2019	3C-Net Narayan et al. (2019)	I3D	40.9	32.3	24.6	-	7.7
	2019	CMCS Liu et al. (2019a)	I3D	41.2	32.1	23.1	15.0	7.0
	2019	MAAN Yuan et al. (2019)	I3D	41.1	30.6	20.3	12.0	6.9
	2019	CleanNet Liu et al. 2019c)	UNT	37.0	30.9	23.9	13.9	7.1
	2020	DGAM Shi et al. (2020)	I3D	46.8	37.5	26.8	17.6	9.0
	2020	TSCN Zhai et al. (2020)	I3D	47.8	37.7	28.7	19.4	10.2
	2021	HAM-Net Islam et al. (2021)	I3D	50.3	41.1	31.0	20.7	11.1
	2021	FAC-Net Huang et al. (2021)	I3D	52.6	44.3	33.4	22.5	12.7
	2022	SFTP Li et al. (2022a)	UNT	47.1	41.7	32.9	22.3	12.5
	2022	A-TSCN Zhai et al. (2022)	I3D	52.1	42.5	33.6	23.4	12.7
	2022	SMEN Sun et al. (2022a)	I3D	60.1	49.4	36.9	23.6	12.9

instance differentiation, drawing closer to instance-level positive pairs in embedded spaces while rejecting negative pairs. In order to bridge the gap between upstream pre-training and downstream tasks, recent contrastive learning approaches have focused on designing excuse tasks specifically for various downstream image tasks, such as object detection (Wang et al. 2021b; Xie et al. 2021; Yang et al. 2021), semantic segmentation (Wang et al. 2021b; Van Gansbeke et al. 2021), etc. In contrast, progress in unsupervised pre-training in respect of video has lagged, and most existing methods (Alwassel et al. 2020; Han et al. 2020; Jenni and Jin 2021; Pan et al. 2021; Qian et al. 2021; Wang et al. 2020) are still designed and evaluated for classified tasks.

Since marking instance actions is tedious and error-prone, precisely demarcating the time boundary of an action instance is time-consuming and subjective for different annotators. The lack of instance-level annotations has inspired recent research on weakly supervised TAD methods. Specifically, there are only rough video-level action categories for each training video rather than labeling them by frame. This represents a new and unexplored area by which to remove labels. The task of temporal action detection in a completely unsupervised environment is called action co-localization (ACL). Only the total number of unique actions occurring in the video dataset is known. Very few studies have attempted this method so far.

Gong et al. (. 2020a) made the first attempt to explore this problem in an unsupervised environment. In order to solve ACL, they proposed a two-step "clustering + localization" iterative process. The clustering step provides the noise pseudo-label for the location step, and the location step provides the time cooperative attention model, thus improving the clustering performance. Using this two-step process, weakly supervised TAD can be viewed as a direct extension of the ACL model. Technically, the authors' contribution is twofold: (1) inspired by the classical image joint segmentation technique (Li et al. 2018a;

Rother et al. 2006), the authors believe that videos of the same action (approximated by the action tag) share a common class-specific collaborative attention model. Temporal collaborative attention models, either class-specific or class-agnostic, learn from video-level tags or pseudo-tags in an iteratively enhanced manner; (2) new losses are specifically designed for ACL, including motion background separation losses and cluster-based triplet losses. Comprehensive evaluations were conducted using 20-action THUMOS14 and 100-action ActivityNet−1.2. The ACL model reached 30.1% (weakly supervised) and 25.0% (unsupervised) for THUMOS14 with an mAP of 0.5.

In addition to directly trying to solve TAD tasks, Zhang et al. (2022) first tried to conduct unsupervised pre-training for TAD tasks. According to their method (Qian et al. 2021), the idea of image contrast learning was applied to the field of video. The authors introduced a time-isovariable contrast-learning paradigm by designing a new unsupervised excuse task, pseudo action localization (PAL). Specifically, the training set was first constructed by cheaply transforming the existing large-scale TAC dataset to simulate TAD custom data with time boundaries. Two temporal regions with random time lengths and proportions were then randomly cropped from one video as pseudo actions. Each region included multiple contiguous segments. They were then pasted to other randomly selected background videos at different time locations. The model can align the pseudo-action features of two synthesised videos with preset time transformations (paste position, clip length, sampling ratio).

7 Conclusion

In this paper, we summarized the literature in respect of temporal action detection published in recent years. Firstly, the background and implementation steps of the video action positioning task were introduced, and then representative and advanced system models introduced in recent years were described according to the working steps. The conclusions are as follows:

(1) All tasks in the field of video understanding first require video feature extraction. In this paper, we divided video feature extraction methods into traditional and deep learning methods and introduced three mainstream deep learning methods;
(2) The second part of the method comprised the core content of this paper: the generation of temporal action proposals. This is used to generate the candidate interval of action to capture the action in the video. We introduced and compared candidate proposal generation for common datasets, including the duration, action type, and the number of actions. Then, evaluation criteria for model algorithms were briefly introduced. Finally, classification and introduction were performed according to anchor-based, boundary-based, and query-based methods;
(3) In terms of learning methods, deep learning can be divided into full supervision, weak supervision, and unsupervised learning. We focused on the classification and summary of weak supervision and provided a comprehensive and detailed introduction to this category.

The goal of video action temporal detection is to detect the action interval and action category in untrimmed video with the highest possible accuracy. The main framework of this article was developed according to the steps of task implementation. According to the

implementation method, the system model was divided into three categories: anchor-based, boundary-based, and query-based. The learning style can be divided into full supervision, weak supervision, and unsupervised learning methods. In this paper, we summarized the performance analyses and development trends; this paper can help scholars to fully understand the temporal action detection task.

8 Prospect

Temporal action detection technology has high application value, and the future development prospects are very broad. Through this review, scholars can further understand and recognize the current development status and trends in this direction. There are still many complicated problems to be solved in the research process. For example, most classifiers have multiple kinds or types and many tunable parameters. The accuracy of most classifiers reported in the literature is not based on extensive analysis of their parameter adjustments. In a paper, one classifier may be misleadingly reported to be less accurate than another, but if its tuning parameters are explored, the classifier's performance may yield better results. In addition, whether the network model is lightweight and questions of quality need to be further explored by researchers. The outlook for the future is as follows:

(1) The production of video datasets should be more biased towards untrimmed videos. On the one hand, this would increase the complexity and authenticity of the dataset; on the other hand, untrimmed video datasets can promote further development of the TAD algorithm;
(2) In terms of model design, more attention should be paid to the weak supervision method in the future to reduce the workload of calibration objects in datasets and improve work efficiency;
(3) The accuracy of most network models has been roughly similar, and the improvement space is small. Despite the high accuracy of models, their multi-module and multi-feature extraction fusion design make them complex, difficult to use, and time-consuming. In order for future models to be more realistic, they should be designed to be more lightweight.

To enhance the relevance of our future outlook, it is essential to reference past related work. Prior research has demonstrated that temporal action detection technology has evolved and made significant progress. For instance, Liu et al. (2022a) proposed an adaptive context model to improve temporal action detection in their research. Furthermore, Wu et al. (2021) also explored the use of Query to introduce Transformer into temporal action detection. These past endeavors offer valuable insights and can guide the trajectory of future developments in the field.

To enhance the relevance of our future outlook, it is essential to reference past related work. Prior research has demonstrated that temporal action detection technology has evolved and made significant progress. For instance,the introduction of the anchor-based method (Shou et al. 2016; Gao et al. 2017a, b; Xu et al. 2017; Lin et al. 2021) from R-CNN has initially solved the problem of action instance segmentation in temporal action detection; and the boundary-based method (Lin et al. 2018; Hsieh et al. 2022; Lin et al. 2019b; Gong et al. 2020b) provides a more flexible action boundary definition for the algorithm framework, enriching the temporal action detection technology. In this

regard, Liu et al. (2022a) proposed a method combining adaptive context modeling to further improve the performance of temporal action detection tasks. At the same time, (Wu et al. 2021) explored how to combine query-based methods with Transformers, further promoting the development of temporal action detection research.

In addition, combining the TAD field with large language models is an interesting and potential research direction. Large-scale language models, such as GPT (Generative Pre-trained Transformer) and Ernie Bot, have achieved great success in the field of natural language processing. These models possess formidable language generation capabilities and can be employed for various natural language processing tasks, including machine translation, text summarization, and question-answering systems. However, training large language models poses considerable challenges, as it necessitates an abundance of text data and extremely powerful computational resources. This makes it difficult for individual users and researchers to train such models from scratch. Consequently, the industry typically adopts the practice of fine-tuning pre-trained large models, which has yielded impressive results in numerous practical application scenarios.

Merging large language models with TAD could result in novel advancements and breakthroughs. TAD can capitalize on the language understanding competencies of large language models to gain a deeper comprehension and interpretation of action content in videos. Additionally, large language models can be employed to generate descriptive action labels or subtitles, thereby enriching video understanding and applications. The TAD area can be meticulously optimized in the following ways:

(1) Action description generation: Equipped with powerful natural language processing capabilities, large language models can create highly accurate, natural, and expressive action descriptions. These descriptions convey the actions in videos with greater detail and precision, enhancing understanding and interpretation of the content.
(2) Contextual understanding and inference: Large language models can perform contextual reasoning and judgment on actions in long videos. By modeling continuous frames or time series, and leveraging the comprehension and reasoning capabilities of the language model, they capture the temporal relationships of actions more effectively.
(3) Action understanding and interpretation: Large language models can provide the ability to interpret and understand detected actions. By analyzing semantic relationships and knowledge bases, combined with action information in videos, more interpretable and understandable descriptions can be generated.

In summary, temporal action detection technology holds great potential for a wide range of future applications. Despite this, the field still confronts numerous complex challenges that need to be addressed, including classifier parameter tuning and the development of lightweight network models. Future advancements can concentrate on creating untrimmed video datasets, implementing weak supervision methods to enhance work efficiency, designing more lightweight models, and fostering the integration of TAD with large language models. Upcoming researchers can build upon past and present accomplishments, exploring innovative avenues to propel the field forward.

Author contributions All authors drafted the manuscript, read, and approved the final manuscript.

Declarations

Conflict of interest The authors declare no competing interests.

References

Abdelgwad M (2021) Arabic aspect based sentiment classification using bert. arXiv: 2107.13290

Abu-El-Haija S, Kothari N, Lee J, et al (2016) Youtube-8m: a large-scale video classification benchmark. arXiv preprint arXiv:1609.08675

Acheampong FA, Nunoo-Mensah H, Chen W (2021) Transformer models for text-based emotion detection: a review of bert-based approaches. Artif Int Rev 54(8):5789–5829

Alwassel H, Mahajan D, Korbar B et al (2020) Self-supervised learning by cross-modal audio-video clustering. Adv Neural Inf Process Syst 33:9758–9770

Alwassel H, Giancola S, Ghanem B (2021) Tsp: temporally-sensitive pretraining of video encoders for localization tasks. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 3173–3183

Arnab A, Dehghani M, Heigold G, et al (2021) Vivit: a video vision transformer. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 6836–6846

Bahdanau D, Cho K, Bengio Y (2014) Neural machine translation by jointly learning to align and translate. Comput Sci. https://doi.org/10.48550/arXiv.1409.0473

Bai Y, Wang Y, Tong Y, et al (2020) Boundary content graph neural network for temporal action proposal generation. In: European conference on computer vision, pp 121–137. Springer

Baraka A, Mohd Noor MH (2022) Weakly-supervised temporal action localization: a survey. Neural Comput Appl 34:1–21

Bertasius G, Wang H, Torresani L (2021) Is space-time attention all you need for video understanding? In: ICML, p 4

Bodla N, Singh B, Chellappa R, et al (2017) Soft-nms–improving object detection with one line of code. In: Proceedings of the IEEE international conference on computer vision, pp 5561–5569

Buch S, Escorcia V, Ghanem B, et al (2019) End-to-end, single-stream temporal action detection in untrimmed videos. In: Proceedings of the British Machine Vision Conference (2019)

Caba Heilbron F, Escorcia V, Ghanem B, et al (2015) Activitynet: A large-scale video benchmark for human activity understanding. In: Proceedings of the ieee conference on computer vision and pattern recognition, pp 961–970

Cao M, Zhang C, Chen L et al (2022) Deep motion prior for weakly-supervised temporal action localization. IEEE Trans Image Process 31:5203–5213

Carion N, Massa F, Synnaeve G, et al (2020) End-to-end object detection with transformers. In: European conference on computer vision, Springer, pp 213–229

Carreira J, Zisserman A (2017) Quo vadis, action recognition? a new model and the kinetics dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6299–6308

Chao YW, Vijayanarasimhan S, Seybold B, et al (2018) Rethinking the faster r-cnn architecture for temporal action localization. In: proceedings of the IEEE conference on computer vision and pattern recognition, pp 1130–1139

Chen C (2004) Searching for intellectual turning points: progressive knowledge domain visualization. Proc Natl Acad Sci 101:5303–5310

Chen C (2006) Citespace ii: detecting and visualizing emerging trends and transient patterns in scientific literature. J Am Soc Inf Sci Technol 57(3):359–377

Chen C (2013) Mapping scientific frontiers: the quest for knowledge visualization. Springer, Berlin

Chen Y, Guo B, Shen Y et al (2021) Boundary graph convolutional network for temporal action detection. Image Vis Comput 109(104):144
Chen C, Ibekwe-SanJuan F, Hou J (2010) The structure and dynamics of Cocitation clusters: a multiple-perspective cocitation analysis. J Am Soc Inf Sci Technol 61(7):1386–1409
Chen J, Ho CM (2022) Mm-vit: Multi-modal video transformer for compressed video action recognition. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 1910–1921
Chen T, Kornblith S, Norouzi M, et al (2020a) A simple framework for contrastive learning of visual representations. In: International conference on machine learning, PMLR, pp 1597–1607
Chen X, Fan H, Girshick R, et al (2020b) Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297
Corona K, Osterdahl K, Collins R, et al (2021) Meva: A large-scale multiview, multimodal video dataset for activity detection. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 1060–1068
Dai Z, Yang Z, Yang Y, et al (2019) Transformer-xl: Attentive language models beyond a fixed-length context. arXiv preprint arXiv:1901.02860
Derrington A, Lennie P (1984) Spatial and temporal contrast sensitivities of Neurones in lateral geniculate nucleus of macaque. J Physiol 357(1):219–240
Devlin J, Chang MW, Lee K, et al (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805
Diba A, Fayyaz M, Sharma V, et al (2017) Temporal 3d convnets: new architecture and transfer learning for video classification. arXiv preprint arXiv:1711.08200
Donahue J, Anne Hendricks L, Guadarrama S, et al (2015) Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2625–2634
Dosovitskiy A, Beyer L, Kolesnikov A, et al (2020) An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929
Escorcia V, Caba Heilbron F, Niebles JC, et al (2016) Daps: Deep action proposals for action understanding. In: European conference on computer vision, Springer, pp 768–784
Van Essen DC, Gallant JL (1994) Neural mechanisms of form and motion processing in the primate visual system. Neuron 13(1):1–10
Fan H, Xiong B, Mangalam K, et al (2021) Multiscale vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 6824–6835
Feichtenhofer C, Pinz A, Wildes RP (2017) Spatiotemporal multiplier networks for video action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4768–4777
Feichtenhofer C, Fan H, Malik J, et al (2019) Slowfast networks for video recognition. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 6202–6211
Felleman DJ, Van Essen DC (1991) Distributed hierarchical processing in the primate cerebral cortex. Cerebral cortex 1(1):1–47
Gao J, Yang Z, Chen K, et al (2017a) Turn tap: Temporal unit regression network for temporal action proposals. In: Proceedings of the IEEE international conference on computer vision, pp 3628–3636
Gao J, Yang Z, Nevatia R (2017b) Cascaded boundary regression for temporal action detection. arXiv preprint arXiv:1705.01180
Gao J, Shi Z, Wang G, et al (2020) Accurate temporal action proposal generation with relation-aware pyramid network. In: Proceedings of the AAAI conference on artificial intelligence, pp 10,810–10,817
Garg S, Vu T, Moschitti A (2020) Tanda: transfer and adapt pre-trained transformer models for answer sentence selection. In: Proceedings of the AAAI conference on artificial intelligence, pp 7780–7788
Ghanem B, Niebles JC, Snoek C, et al (2017) Activitynet challenge 2017 summary. arXiv preprint arXiv:1710.08011
Ghorbani S, Mahdaviani K, Thaler A, et al (2020) Movi: a large multipurpose motion and video dataset. arXiv preprint arXiv:2003.01888
Girdhar R, Carreira J, Doersch C, et al (2019) Video action transformer network. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 244–253
Girshick R, Donahue J, Darrell T, et al (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 580–587
Gong G, Wang X, Mu Y, et al (2020a) Learning temporal co-attention models for unsupervised video action localization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9819–9828

Gong G, Zheng L, Mu Y (2020b) Scale matters: Temporal scale aggregation network for precise action localization in untrimmed videos. In: 2020 IEEE international conference on multimedia and expo (ICME), IEEE, pp 1–6
Gorelick L, Blank M, Shechtman E et al (2007) Actions as space-time shapes. IEEE Trans Pattern Anal Mach Intell 29(12):2247–2253
Graziani M, Dutkiewicz L, Calvaresi D et al (2022) A global taxonomy of interpretable AI: unifying the terminology for the technical and social sciences. Artif Intell Rev 56:1–32
Grill JB, Strub F, Altché F et al (2020) Bootstrap your own latent-a new approach to self-supervised learning. Adv Neural Inf Process Syst 33:21,271-21,284
Gu C, Sun C, Ross DA, et al (2018) Ava: A video dataset of spatio-temporally localized atomic visual actions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6047–6056
Guo H, Wang H, Ji Q (2022) Uncertainty-guided probabilistic transformer for complex action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 20,052–20,061
Gutmann M, Hyvärinen A (2010) Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In: Proceedings of the thirteenth international conference on artificial intelligence and statistics, JMLR Workshop and Conference Proceedings, pp 297–304
Han T, Xie W, Zisserman A (2020) Self-supervised co-training for video representation learning. Adv Neural Inf Process Syst 33:5679–5690
He K, Fan H, Wu Y, et al (2020) Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9729–9738
Heilbron FC, Niebles JC, Ghanem B (2016) Fast temporal activity proposals for efficient detection of human actions in untrimmed videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1914–1923
Horn G, Perona P (2017) The devil is in the tails: fine-grained classification in the wild. arXiv preprint arXiv:1709.01450 2
Hsieh HY, Chen DJ, Liu TL (2022) Contextual proposal network for action localization. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 2129–2138
Hu K, Ding Y, Jin J et al (2022) Skeleton motion recognition based on multi-scale deep Spatio-temporal features. Appl Sci 12(3):1028
Hu K, Ding Y, Jin J et al (2022) Multiple attention mechanism graph convolution HAR model based on coordination theory. Sensors 22(14):5259
Hu K, Jin J, Zheng F et al (2022) Overview of behavior recognition based on deep learning. Artif Intell Rev 56:1–33
Hu K, Weng C, Shen C et al (2023) A multi-stage underwater image aesthetic enhancement algorithm based on a generative adversarial network. Eng Appl Artif Intell 123(106):196
Hu K, Zheng F, Weng L et al (2021) Action recognition algorithm of Spatio-temporal differential LSTM based on feature enhancement. Appl Sci 11(17):7876
Huang L, Huang Y, Ouyang W, et al (2020) Relational prototypical network for weakly supervised temporal action localization. In: proceedings of the AAAI conference on artificial intelligence, pp 11,053–11,060
Huang L, Wang L, Li H (2021) Foreground-action consistency network for weakly supervised temporal action localization. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 8002–8011
Hubel DH, Wiesel TN (1965) Receptive fields and functional architecture in two Nonstriate visual areas (18 and 19) of the cat. J Neurophys 28(2):229–289
Hutchinson MS, Gadepally VN (2021) Video action understanding: a tutorial. IEEE Access 9:134611–134637
Islam A, Long C, Radke R (2021) A hybrid attention mechanism for weakly-supervised temporal action localization. In: Proceedings of the AAAI conference on artificial intelligence, pp 1637–1645
Jenni S, Jin H (2021) Time-equivariant contrastive video representation learning. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 9970–9980
Jia Y, Shelhamer E, Donahue J, et al (2014) Caffe: Convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM international conference on multimedia, pp 675–678
Jiang YG, Liu J, Roshan Zamir A, et al (2014) THUMOS challenge: action recognition with a large number of classes. http://crcv.ucf.edu/THUMOS14/
Ke Y, Sukthankar R, Hebert M (2007) Event detection in crowded videos. In: 2007 IEEE 11th international conference on computer vision, IEEE, pp 1–8

Klaser A, Marszałek M, Schmid C (2008) A spatio-temporal descriptor based on 3d-gradients. In: BMVC 2008-19th British machine vision conference, British Machine Vision Association, pp 275–1
Kong Y, Fu Y (2022) Human action recognition and prediction: a survey. Int J Comput Vis 130(5):1366–1401
Kotsiantis SB, Zaharakis ID, Pintelas PE (2006) Machine learning: a review of classification and combining techniques. Artif Intell Rev 26(3):159–190
Krizhevsky A, Sutskever I, Hinton GE (2017) Imagenet classification with deep convolutional neural networks. Commun ACM 60(6):84–90
Kuehne H, Jhuang H, Garrote E, et al (2011) Hmdb: a large video database for human motion recognition. In: 2011 International conference on computer vision, IEEE, pp 2556–2563
Kumar S, Garg S, Mehta K, et al (2019) Improving answer selection and answer triggering using hard negatives. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), pp 5911–5917
Kumar Singh K, Jae Lee Y (2017) Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In: Proceedings of the IEEE international conference on computer vision, pp 3524–3533
Lan Z, Chen M, Goodman S, et al (2019) Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942
Laptev I (2005) On space-time interest points. Int J Comput Vis 64(2):107–123
Lauriola I, Lavelli A, Aiolli F (2022) An introduction to deep learning in natural language processing: models, techniques, and tools. Neurocomputing 470:443–456
Le N, Rathour VS, Yamazaki K et al (2021) Deep reinforcement learning in computer vision: a comprehensive survey. Artif Intell Rev 55:1–87
Lee J, Yoon W, Kim S et al (2020) Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36(4):1234–1240
Lee P, Uh Y, Byun H (2020b) Background suppression network for weakly-supervised temporal action localization. In: Proceedings of the AAAI conference on artificial intelligence, pp 11,320–11,327
Li Z, Gavrilyuk K, Gavves E et al (2018) Videolstm convolves, attends and flows for action recognition. Comput Vis Image Understand 166:41–50
Li B, Guo B, Zhu Y et al (2022) Superframe-based temporal proposals for weakly supervised temporal action detection. IEEE Trans Multimed. https://doi.org/10.1109/TMM.2022.3163459
Li M, Huang B, Tian G (2022) A comprehensive survey on 3d face recognition methods. Engineering Applications of Artificial Intelligence 110(104):669
Li W, Hosseini Jafari O, Rother C (2018a) Deep object co-segmentation. In: Asian conference on computer vision, Springer, pp 638–653
Li X, Wang W, Hu X, et al (2019) Selective kernel networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 510–519
Li Y, Lin W, See J, et al (2020) Cfad: Coarse-to-fine action detector for spatiotemporal action localization. In: European conference on computer vision, Springer, pp 510–527
Lin C, Li J, Wang Y, et al (2020) Fast learning of temporal action proposal via dense boundary generator. In: Proceedings of the AAAI conference on artificial intelligence, pp 11,499–11,506
Lin C, Xu C, Luo D, et al (2021) Learning salient boundary feature for anchor-free temporal action localization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3320–3329
Lin J, Gan C, Han S (2019a) Tsm: Temporal shift module for efficient video understanding. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 7083–7093
Lin T, Zhao X, Shou Z (2017) Single shot temporal action detection. In: Proceedings of the 25th ACM international conference on Multimedia, pp 988–996
Lin T, Zhao X, Su H, et al (2018) Bsn: Boundary sensitive network for temporal action proposal generation. In: Proceedings of the European conference on computer vision (ECCV), pp 3–19
Lin T, Liu X, Li X, et al (2019b) Bmn: Boundary-matching network for temporal action proposal generation. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 3889–3898
Liu X, Wang Q, Hu Y et al (2022) End-to-end temporal action detection with transformer. IEEE Trans Image Process 31:5427–5441
Liu Y, Wang L, Wang Y et al (2022) Fineaction: a fine-grained video dataset for temporal action localization. IEEE Trans Image Process 31:6937–6950
Liu Q, Wang Z (2020) Progressive boundary refinement network for temporal action detection. In: Proceedings of the AAAI conference on artificial intelligence, pp 11,612–11,619

Liu W, Anguelov D, Erhan D, et al (2016) Ssd: Single shot multibox detector. In: European conference on computer vision, Springer, pp 21–37
Liu Y, Ott M, Goyal N, et al (2019b) Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692
Liu Z, Wang L, Zhang Q, et al (2019c) Weakly supervised temporal action localization through contrast based evaluation networks. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 3899–3908
Liu Z, Lin Y, Cao Y, et al (2021) Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 10,012–10,022
Liu D, Jiang T, Wang Y (2019a) Completeness modeling and context separation for weakly supervised temporal action localization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 1298–1307
Liu J, Luo J, Shah M (2009) Recognizing realistic actions from videos "in the wild". In: 2009 IEEE conference on computer vision and pattern recognition, IEEE, pp 1996–2003
Livingstone M, Hubel D (1988) Segregation of form, color, movement, and depth: anatomy, physiology, and perception. Science 240(4853):740–749
Long F, Yao T, Qiu Z, et al (2019) Gaussian temporal awareness networks for action localization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 344–353
Majd M, Safabakhsh R (2020) Correlational convolutional LSTM for human action recognition. Neurocomputing 396:224–229
Marszalek M, Laptev I, Schmid C (2009) Actions in context. In: 2009 IEEE conference on computer vision and pattern recognition, IEEE, pp 2929–2936
Monfort M, Andonian A, Zhou B et al (2019) Moments in time dataset: one million videos for event understanding. IEEE Trans Pattern Anal Mach Intell 42(2):502–508
Muhammad K, Ullah A, Imran AS et al (2021) Human action recognition using attention based LSTM network with dilated CNN features. Future Gener Comput Syst 125:820–830
Narayan S, Cholakkal H, Khan FS, et al (2019) 3c-net: Category count and center loss for weakly-supervised action localization. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 8679–8687
Narkhede MV, Bartakke PP, Sutaone MS (2022) A review on weight initialization strategies for neural networks. Artif intell Rev 55(1):291–322
Nguyen P, Liu T, Prasad G, et al (2018) Weakly supervised action localization by sparse temporal pooling network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6752–6761
Niebles JC, Chen CW, Fei-Fei L (2010) Modeling temporal structure of decomposable motion segments for activity classification. In: European conference on computer vision, Springer, pp 392–405
Ning R, Zhang C, Zou Y (2021) 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, pp 2460–2464
Oneata D, Verbeek J, Schmid C (2014a) Efficient action localization with approximately normalized fisher vectors. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2545–2552
Oneata D, Verbeek J, Schmid C (2014b) The lear submission at thumos 2014. ECCV THUMOS Workshop (2014)
Oord Avd, Li Y, Vinyals O (2018) Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748
Pan T, Song Y, Yang T, et al (2021) Videomoco: Contrastive video representation learning with temporally adversarial examples. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 11,205–11,214
Pareek P, Thakkar A (2021) A survey on video-based human action recognition: recent updates, datasets, challenges, and applications. Artif Intell Rev 54(3):2259–2322
Patrick M, Campbell D, Asano Y et al (2021) Keeping your eye on the ball: trajectory attention in video transformers. Adv Neural Inf Process Syst 34:12,493-12,506
Paul S, Roy S, Roy-Chowdhury AK (2018) W-talc: Weakly-supervised temporal activity localization and classification. In: Proceedings of the European conference on computer vision (ECCV), pp 563–579
Qian R, Meng T, Gong B, et al (2021) Spatiotemporal contrastive video representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 6964–6974
Qing Z, Su H, Gan W, et al (2021) Temporal context aggregation network for temporal action proposal refinement. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 485–494

Qiu Z, Yao T, Mei T (2017) Learning spatio-temporal representation with pseudo-3d residual networks. In: proceedings of the IEEE international conference on computer vision, pp 5533–5541
Radford A, Wu J, Child R et al (2019) Language models are unsupervised multitask learners. OpenAI blog 1(8):9
Radford A, Narasimhan K, Salimans T, et al (2018) Improving language understanding by generative pre-training. OpenAI blog, 2018.
Reddy KK, Shah M (2013) Recognizing 50 human action categories of web videos. Mach Vis Appl 24(5):971–981
Redmon J, Divvala S, Girshick R, et al (2016) You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 779–788
Ren S, He K, Girshick R, et al (2015) Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 28:91–99, 2015.
Rodriguez MD, Ahmed J, Shah M (2008) Action mach a spatio-temporal maximum average correlation height filter for action recognition. In: 2008 IEEE conference on computer vision and pattern recognition, IEEE, pp 1–8
Rother C, Minka T, Blake A, et al (2006) Cosegmentation of image pairs by histogram matching-incorporating a global constraint into mrfs. In: 2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06), IEEE, pp 993–1000
Ruan L, Jin Q (2022) Survey: transformer based video-language pre-training. AI Open 3:1–13
Sadanand S, Corso JJ (2012) Action bank: A high-level representation of activity in video. In: 2012 IEEE Conference on computer vision and pattern recognition, IEEE, pp 1234–1241
Sanh V, Debut L, Chaumond J, et al (2019) Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108
Satkin S, Hebert M (2010) Modeling the temporal extent of actions. In: European conference on computer vision, Springer, pp 536–548
Schuldt C, Laptev I, Caputo B (2004) Recognizing human actions: a local svm approach. In: Proceedings of the 17th international conference on pattern recognition, 2004. ICPR 2004., IEEE, pp 32–36
Scovanner P, Ali S, Shah M (2007) A 3-dimensional sift descriptor and its application to action recognition. In: Proceedings of the 15th ACM international conference on Multimedia, pp 357–360
Shao T, Guo Y, Chen H et al (2019) Transformer-based neural network for answer selection in question answering. IEEE Access 7:26146–26156
Sharma S, Kiros R, Salakhutdinov R (2015) Action recognition using visual attention. arXiv preprint arXiv:1511.04119
Shi B, Dai Q, Mu Y, et al (2020) Weakly-supervised action localization by generative attention modeling. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 1009–1019
Shi D, Zhong Y, Cao Q, et al (2022) React: Temporal action detection with relational queries. In: European conference on computer vision, Springer, pp 105–121
Shi X, Chen Z, Wang H, et al (2015) Convolutional lstm network: A machine learning approach for precipitation nowcasting. In Advances in neural information processing systems, 802–810.
Shou Z, Wang D, Chang SF (2016) Temporal action localization in untrimmed videos via multi-stage cnns. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1049–1058
Shou Z, Gao H, Zhang L, et al (2018) Autoloc: Weakly-supervised temporal action localization in untrimmed videos. In: Proceedings of the European conference on computer vision (ECCV), pp 154–171
Sigurdsson GA, Varol G, Wang X, et al (2016) Hollywood in homes: Crowdsourcing data collection for activity understanding. In: European conference on computer vision, Springer, pp 510–526
Simonyan K, Zisserman A (2014) Two-stream convolutional networks for action recognition in videos. Advances in neural information processing systems, 27, 2014.
Singh KK, Xiao F, Lee YJ (2016) Track and transfer: Watching videos to simulate strong human supervision for weakly-supervised object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3548–3556
HO, Lee YJ, Jegelka S, et al (2014) Weakly-supervised discovery of visual pattern configurations. In Advances in Neural Information Processing Systems, pages 1637–1645, 2014.
Song Y, Vallmitjana J, Stent A, et al (2015) Tvsum: Summarizing web videos using titles. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5179–5187
Soomro K, Zamir AR, Shah M (2012) Ucf101: a dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402
Srivastava N, Mansimov E, Salakhudinov R (2015) Unsupervised learning of video representations using lstms. In: International conference on machine learning, PMLR, pp 843–852

Sudhakaran S, Escalera S, Lanz O (2019) Lsta: Long short-term attention for egocentric action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9954–9963
Sun Z, Ke Q, Rahmani H et al (2022) Human action recognition from various data modalities: a review. IEEE Trans Pattern Anal Mach Intell. https://doi.org/10.1109/TPAMI.2022.3183112
Sun W, Su R, Yu Q et al (2022) Slow motion matters: a slow motion enhanced network for weakly supervised temporal action localization. IEEE Trans Circ Syst Video Technol 33(1):354–366
Sun C, Qiu X, Xu Y, et al (2019) How to fine-tune bert for text classification? In: China national conference on Chinese computational linguistics, Springer, pp 194–206
Tan J, Tang J, Wang L, et al (2021) Relaxed transformer decoders for direct action proposal generation. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 13,526–13,535
Tian Z, Shen C, Chen H, et al (2019) Fcos: Fully convolutional one-stage object detection. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 9627–9636
Tran D, Bourdev L, Fergus R, et al (2015) Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE international conference on computer vision, pp 4489–4497
Tran D, Wang H, Torresani L, et al (2018) A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6450–6459
Truong TD, Bui QH, Duong CN, et al (2022) Direcformer: A directed attention in transformer approach to robust action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 20,030–20,040
Vahdani E, Tian Y (2022) Deep learning-based action detection in untrimmed videos: a survey. IEEE Trans Pattern Anal Mach Intell. https://doi.org/10.1109/TPAMI.2022.3193611
Van Gansbeke W, Vandenhende S, Georgoulis S, et al (2021) Unsupervised semantic segmentation by contrasting object mask proposals. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 10,052–10,062
Varol G, Laptev I, Schmid C (2017) Long-term temporal convolutions for action recognition. IEEE Trans Pattern Anal Mach Intell 40(6):1510–1517
Vaswani A, Shazeer N, Parmar N, et al (2017) Attention is all you need. In Advances in neural information processing systems, pp. 5998–6008, 2017.
Wang H, Kläser A, Schmid C et al (2013) Dense trajectories and motion boundary descriptors for action recognition. Int J Comput Vis 103(1):60–79
Wang Z, Lu H, Jin J et al (2022) Human action recognition based on improved two-stream convolution network. Appl Sci 12(12):5784
Wang H, Wu H, He Z et al (2021) Progress in machine translation. Engineering 18:143–153
Wang H, Schmid C (2013) Action recognition with improved trajectories. In: Proceedings of the IEEE international conference on computer vision, pp 3551–3558
Wang J, Jiao J, Liu YH (2020) Self-supervised video representation learning by pace prediction. In: European conference on computer vision, Springer, pp 504–521
Wang L, Qiao Y, Tang X (2014) Video action detection with relational dynamic-poselets. In: European conference on computer vision, Springer, pp 565–580
Wang L, Qiao Y, Tang X, et al (2016a) Actionness estimation using hybrid fully convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2708–2717
Wang L, Xiong Y, Wang Z, et al (2016b) Temporal segment networks: towards good practices for deep action recognition. In: European conference on computer vision, Springer, pp 20–36
Wang L, Xiong Y, Lin D, et al (2017) Untrimmednets for weakly supervised action recognition and detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4325–4334
Wang Q, Zhang Y, Zheng Y, et al (2022a) Rcl: recurrent continuous localization for temporal action detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 13,566–13,575
Wang X, Girshick R, Gupta A, et al (2018) Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7794–7803
Wang X, Zhang R, Shen C, et al (2021b) Dense contrastive learning for self-supervised visual pre-training. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3024–3033
Weinland D, Boyer E, Ronfard R (2007) Action recognition from arbitrary views using 3d exemplars. In: 2007 IEEE 11th international conference on computer vision, IEEE, pp 1–7

Wu J, Sun P, Chen S, et al (2021) Towards high-quality temporal action detection with sparse proposals. arXiv preprint arXiv:2109.08847
Wu Z, Xiong C, Jiang YG, et al (2019) Liteeval: a coarse-to-fine framework for resource efficient video recognition. In Advances in Neural Information Processing Systems, 7778–7787.
Xia H, Zhan Y (2020) A survey on temporal action localization. IEEE Access 8:70477–70487
Xie E, Ding J, Wang W, et al (2021) Detco: unsupervised contrastive learning for object detection. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 8392–8401
Xiong Y, Zhao Y, Wang L, et al (2017) A pursuit of temporal accuracy in general activity detection. arXiv preprint arXiv:1703.02716
Xu M, Perez Rua JM, Zhu X et al (2021) Low-fidelity video encoder optimization for temporal action localization. Adv Neural Inf Process Syst 34:9923–9935
Xu H, Das A, Saenko K (2017) R-c3d: Region convolutional 3d network for temporal activity detection. In: Proceedings of the IEEE international conference on computer vision, pp 5783–5792
Xu M, Pérez-Rúa JM, Escorcia V, et al (2021a) Boundary-sensitive pre-training for temporal localization in videos. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 7220–7230
Xu Y, Zhang C, Cheng Z, et al (2019) Segregated temporal assembly recurrent networks for weakly supervised multiple action detection. In: Proceedings of the AAAI conference on artificial intelligence, pp 9070–9078
Yadav A, Vishwakarma DK (2020) Sentiment analysis using deep learning architectures: a review. Artif Intell Rev 53(6):4335–4385
Yang L, Peng H, Zhang D et al (2020) Revisiting anchor mechanisms for temporal action localization. IEEE Trans Image Process 29:8535–8548
Yang C, Wu Z, Zhou B, et al (2021) Instance localization for self-supervised detection pretraining. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3987–3996
Yeung S, Russakovsky O, Mori G, et al (2016) End-to-end learning of action detection from frame glimpses in videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2678–2687
Yu G, Yuan J (2015) Fast action proposals for human action detection and search. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1302–1311
Yu X, Hu W, Lu S, et al (2019) Biobert based named entity recognition in electronic medical record. In: 2019 10th international conference on information technology in medicine and education (ITME), IEEE, pp 49–52
Yuan J, Ni B, Yang X, et al (2016) Temporal action localization with pyramid of score distribution features. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3093–3102
Yuan Y, Lyu Y, Shen X, et al (2019) Marginalized average attentional network for weakly-supervised learning. arXiv preprint arXiv:1905.08586
Yue-Hei Ng J, Hausknecht M, Vijayanarasimhan S, et al (2015) Beyond short snippets: Deep networks for video classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4694–4702
Zha X, Zhu W, Xun L et al (2021) Shifted chunk transformer for Spatio-temporal representational learning. Adv Neural Inf Process Syst 34:11384–11396
Zhai Y, Wang L, Tang W et al (2022) Adaptive two-stream consensus network for weakly-supervised temporal action localization. IEEE Trans Pattern Anal Mach Intell 45(4):4136–4151
Zhai Y, Wang L, Tang W, et al (2020) Two-stream consensus network for weakly-supervised temporal action localization. In: European conference on computer vision, Springer, pp 37–54
Zhang Z, Tao D (2012) Slow feature analysis for human action recognition. IEEE Trans Pattern Anal Mach Intell 34(3):436–450
Zhang C, Yang T, Weng J, et al (2022) Unsupervised pre-training for temporal action localization tasks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 14,031–14,041
Zhao Y, Xiong Y, Wang L, et al (2017) Temporal action detection with structured segment networks. In: Proceedings of the IEEE international conference on computer vision, pp 2914–2923

Authors and Affiliations

Kai Hu[1,2] · Chaowen Shen[1,2] · Tianyan Wang[1,2] · Keer Xu[1,2] · Qingfeng Xia[3,4] · Min Xia[1,2] · Chengxue Cai[1,2]

✉ Kai Hu
001600@nuist.edu.cn

Chaowen Shen
20211249071@nuist.edu.cn

Tianyan Wang
20211249158@nuist.edu.cn

Keer Xu
202212490502@nuist.edu.cn

Qingfeng Xia
xqf@cwxu.edu.cn

Min Xia
xiamin@nuist.edu.cn

Chengxue Cai
202183240085@nuist.edu.cn

[1] School of Automation, Nanjing University of Information Science and Technology, No.219, Ningliu Road, Nanjing 210044, Jiangsu, China

[2] Jiangsu Collaborative Innovation Center of Atmospheric Environment and Equipment Technology (CICAEET), Nanjing University of Information Science and Technology, No.219, Ningliu Road, Nanjing 210044, Jiangsu, China

[3] School of Automation, Wuxi University, No.333, Xishan Road, Wuxi 214105, Jiangsu, China

[4] School of Management and Engineering, Nanjing University, No.22, Hankou Rd, Gulou District, Nanjing 210093, Jiangsu, China