IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 19, 2020, accepted March 30, 2020, date of publication April 9, 2020, date of current version April 27, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2986861

A Survey on Temporal Action Localization

HUIFEN XIA[1,2] AND YONGZHAO ZHAN[1]
[1]Department of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang 212013, China
[2]Changzhou Vocational Institute of Mechatronic Technology, Changzhou 213164, China

Corresponding author: Yongzhao Zhan (yzzhan@ujs.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61672268.

ABSTRACT Temporal action localization is one of the most crucial and challenging problems for video understanding in computer vision. It has received a lot of attention in recent years because of the extensive application of daily life. Temporal action localization has made some significant progress, especially with the development of deep learning recently. And more demand is for temporal action localization in untrimmed videos. In this paper, our target is to survey the state-of–the-art techniques and models for video temporal action localization. It mainly includes the related techniques, some benchmark datasets and the evaluation metrics of temporal action localization. In addition, we summarize temporal action localization from two aspects: fully-supervised learning and weakly-supervised learning. And we list several representative works and compare their performances respectively. Finally, we make some deep analysis and propose potential research directions, and conclude the survey.

INDEX TERMS Action detection, computer vision, fully-supervised learning, temporal action localization, weakly-supervised learning.

I. INTRODUCTION

With the number of videos grows tremendously, video understanding becomes a hot question and a challenging direction in computer vision. The video understanding direction includes many sub-research directions. According to ActivityNet Challenge 2017 [48] held by CVPR in Hawaii, a total of 5 tasks were proposed.

a) Untrimmed Video Classification (ActivityNet [7]).

b) Trimmed Action Recognition (Kinetics [44]).

c) Temporal Action Proposals (ActivityNet).

d) Temporal Action Localization (ActivityNet).

e) Dense-Captioning Events in Videos (ActivityNet Captions).

In this survey, we focus on temporal action localization, which is the 4th of the above lists. It requires the detections of temporal intervals which contain the target actions. For a long untrimmed video, temporal action localization mainly solves two tasks which are recognition and localization. Specifically, a) When does the action occur, that is the start time and the end time of the action. b) What category does each proposal belong to (such as Waving, Climbing, or Basketball-Dunk). Of course, a video may contain one or more action clips. So temporal action localization is to develop models and techniques which provide the most basic information needed by computer vision applications: What are the actions and when do the actions happen? We take this task as action localization, or temporal action localization, or action detection.

Although both action recognition and action localization are important tasks of video understanding, temporal action localization is more challenging than action recognition. And the relationship between action recognition and action localization is similar to image recognition and image detection. But owning to the temporal series information, temporal action localization is much difficult than image detection. The difficulties are as follows: a) temporal information. Because of the 1-dimension temporal series information, temporal action localization can't use static image information. It must combine the information of temporal series. b) Unclear boundaries. Different from object detection, the boundaries of the object are usually very clear, so we can mark out a clearer bounding box for the object. However, there might have not sensible definition about the exact temporal extent of action [56], [57]. So it's impossible to give an accurate boundary when the action starts and when the action ends. c) Large temporal spans. The span of temporal action fragments can be very large. For example, waving hands can only take a few seconds while climbing or cycling can last for tens of minutes. Their spans differ in length which make them extremely difficult to extract proposals. In addition, in the

The associate editor coordinating the review of this manuscript and approving it for publication was Long Wang.

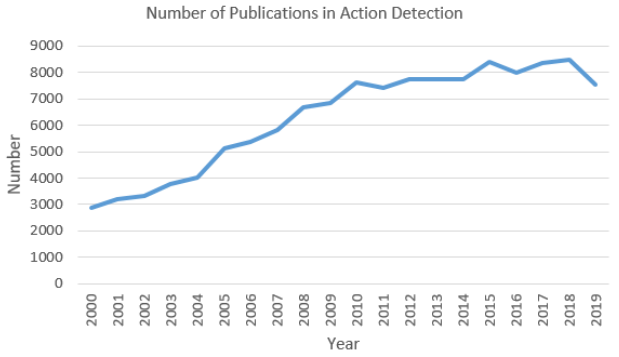


FIGURE 1. The increasing number of publications of academic and conference papers in action detection from 2000 to 2019. (Some declines in 2019 are due to incomplete statistics. Data is from Superstar Discovery advanced search).

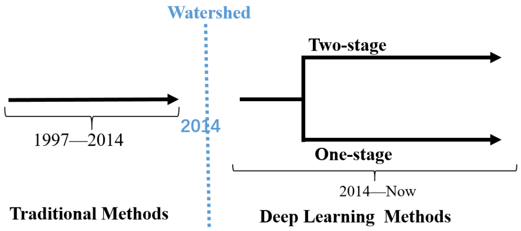


FIGURE 2. The timeline frame of temporal action localization.

open environment, there are many problems such as multi-scale, multi-target, and camera movement.

Temporal action localization is very close to our life, it has extensive application prospects and social value in the fields of video summarization [51], public video surveillance [49], skill assessment [50] and daily life security. So it has received a lot of attention in recent years. The total number of publications related to ''action detection'' is about 324,127, including books, journals, dissertations, conference papers, patents and some scientific and technological achievements in the past twenty years. Below we mainly analyze the number of publications trends of academic and conference papers on action detection, which is shown as FIGURE 1.

This survey is intended to help the beginners who are interested in temporal action localization. It provides an overview of methods and recent developments of action localization. The rest of this article is organized as follows. Section II outlines the related techniques. Section III introduces the benchmark datasets for temporal action localization. Section IV describes the performance evaluation metrics of the models. Section V provides an overview of action localization models and methods from fully-supervised and weakly-supervised learning. Section VI discusses the current challenges and suggests future directions. Section VII conclude the paper.

II. RELATED TECHNIQUES

Since temporal action localization has been an active research area recently, many different approaches to deal with this problem have come into being. Although action detection has been studied for many years, it is still in the test phase of laboratory datasets, and there is no actual practicality and industrialization. The task of understanding what and when the action happens in a video is very challenging. It can be seen that there is still no robust solution for this task currently. In this section, we will review the related techniques of temporal action localization.

As we know, video feature representation can provide useful information for video action and a lot of attempts have been made. In the past twenty years, it's well known that the progress of feature extraction has generally gone through two important historical periods. One is the traditional action detection periods before 2014, the other is deep learning based periods after 2014. The timeline frame is shown as FIGURE 2.

In deep learning periods, they are mainly grouped into two type frameworks: ''two-stage detection'' and ''one-stage detection''. Specifically, the former is based on the ''proposal-then-classification'' paradigm which is the mainstream method. The latter does proposal and classification simultaneously, so we call it one-stage detection.

A. TRADITIONAL METHODS

Due to action recognition is a part of temporal action localization, so most of the early action localization algorithms were built based on hand-crafted features, the same as action recognition. There are several ways to extract video features that contain static image features and temporal visual features. Speaking specifically, static image features are SIFT (Scale-Invariant Feature Transform) [62], [63] and HOG (Histogram of Oriented Gradients) [64] and so on. HOG can be considered as an improvement of SIFT. While temporal visual features are the combination of static image features and temporal information. By these features, temporal information of a video can be achieved.

In general, we can divide feature extraction into local feature extraction and global feature extraction. a) Local features extraction refers to local points of interest or regions of interest in the video. It includes statistics, dictionary learning, bag-of-words (BoW) [65], [66] and feature learning and so on. Compared with global features, local features are more robust to video lighting, perspective, camera shake, and complex backgrounds. b) Global features extraction refers to the overall characteristics of human behavior, such as the contours and skeleton of the human body. It includes global density and trajectory methods. To solve the problem of human behavior in complex scenes, it's not enough just to detect the gray level changes in the spatiotemporal area. Therefore, researchers have proposed many feature extraction methods based on feature point tracking. The approximate process is as follows: These methods detect feature points in the temporal region of the video firstly, then track these feature points frame by frame and join the trajectories that form the feature points. At last, they use feature descriptors to describe the trajectory and its temporal neighborhood. Among the many feature extraction methods based on feature point tracking, the classic method is Dense Trajectories (DT) [67]. Afterward, considering that the camera movement leads to the extraction of DT features which are not related to human behavior, the DT feature is further improved, and an improved dense trajectory (iDT) [1] method is proposed. Although many today's methods have far surpassed iDT, the valuable insight of iDT is still influenced for the later research work. It is worth noting that the combination of deep learning and iDT can usually further improve the performance. Many papers have adopted the form of "Our method + iDT" to achieve the highest level SOTA (state-of-the-Art).

In any case, the research process and ideas of traditional feature extraction methods are very useful because these methods have strong interpretability. They provide inspiration and analogy for designing deep learning methods to solve such problems.

B. DEEP LEARNING METHODS

As the performance of the methods using hand-crafted features became stabilized, temporal action localization has reached a plateau. Along with the convolutional neural network was rebirthed [61], a lot of works have risen. The convolutional neural network can learn the robust and high-level feature representations. For example, a 2D-CNN [2] for large-scale video (Sports-1M dataset) classification was proposed of Li Feifei's group in 2014. Although its performance hasn't compared to the methods based on hand-crafted features, the idea inspired later research. Afterwards, two-stream CNNs [3] (RGB frames and optical flow), 3D convolutional networks [4] and then their variations become the popular solutions to learn discriminative features for action recognition. Subsequently, a combination of two-stream and C3D networks named as I3D [5] (Inception 3D) was proposed. And it has become a generic video feature representation encoder. In addition, several methods based on recurrent neural networks [53], [54] were introduced to capture the long range dynamics for action recognition. TSN [55] was designed to model the entire video information with average aggregation via the strategy of a sparse sampling. According to FIGURE 1, we know deep learning is divided into two types: two-stage localization and one-stage localization.

1) TWO-STAGE LOCALIZATION METHODS

Two-stage type is based on the paradigm of proposal-then-classification. This paradigm extracts temporal proposals first, and then deals with the classification and regression operation. This type is the mainstream method, so most papers adopt this method. In fact, the generation of proposals is a difficult point in this paradigm for temporal action localization which is similar to proposal generation in object detection (region proposals generation in R-CNN [68]). A good proposal algorithm can greatly improve the effect of the model.

The task of temporal action proposal generation is to generate a certain number of temporal proposals for an untrimmed long video. A temporal action proposal is a temporal interval that may contain action segments (from the start boundary to the end boundary). Generally, average recall (AR) under a certain number of proposals is used to measure the performance of algorithm. The datasets used commonly are THUMOS14 [6] and ActivityNet [7], There are several methods for extracting proposals.

a: SLIDING WINDOW (S-CNN [14], 2016)

In 2016, S-CNN is the first method to fix some size sliding windows to generate various sizes video segments, and then deal with them by a multi-stage network (Segment-CNN). SCNN includes three sub-networks all using C3D network. The first one is proposal network which is used to determine the probability of the current segment being an action. The second one is classification network which is used to classify video segments. The third one is localization networks whose output is still the probability of the category. And an overlap-related loss function is added during training so that the network can estimate the category and overlap of a video clip better. In principle, when the degree of overlap is higher, the effect is better. But the amount of calculation is very large. Finally, non-maximized suppression (NMS) is used to remove overlapping segments and complete the prediction.

Theoretically, this method is the most comprehensive as long as the overlap is high enough, but it has more redundancy.

b: TEMPORAL ACTIONNESS GROUPING (TAG [15], 2017)

Previous work used sliding windows to extract proposals, but this method cannot deal with video actions of different lengths. Because in general action recognition, convolution is applied to dense video frames. And the consumption is so huge for long action videos.

Y. Xiong *et al.* [15] proposes a new framework to accurately determine the boundaries of action for variable-length videos in 2017. The framework contains two parts: generating temporal proposals and classifying proposed candidates. The former generates s series of proposals, and the latter determines whether it is an action and predicts its category. In order to generate a temporal proposal, TAG network is proposed. They are three main steps: a) Extract snippets: Each snippet contains a video frame and optical flow information, and snippets are obtained at regular intervals. b) Actionness: Determine whether a snippet contains actions. In order to do it, it learns a binary classification network using TSN (Temporal Segment Network). c) Grouping: For the output snippets sequences and their probabilities, it will group those continuous snippets with high scores. At the same time, setting some thresholds to remove those snippets with lower scores for preventing noise interference, and generally setting multiple sets of thresholds to prevent missing proposals.

This method is more flexible for boundaries, but it may miss some proposals due to classification errors.

c: TEMPORAL UNIT REGRESS NETWORK (TURN TAP [16], 2017)

In SCNN network, it used sliding windows to find the proposals. If you want to get accurate results, you need to increase the overlap between the windows which results in a problem of large calculation.

In order to reduce the amount of calculation and increase the accuracy of temporal localization, Gao J.Y. *et al.* [16] proposed TURN learning from the method of boundary regression introduced by faster-RCNN [69], [70] in 2017. This method divides the video into fixed size units, such as a unit of 16 video frames, then puts each unit into C3D to extract the horizontal features. Adjacent units form a clip and let each unit as an anchor unit construct a clip pyramid. Then temporal coordinate regression is performed at the unit. The network contains two outputs: The first output is confidence score that determines whether the clip contains actions; and the second output is temporal coordinates offset which adjusts the boundary.

The main contributions are as follows: (1) A novel method for generating temporal proposal segments using coordinate regression. (2) Fast speed (800fps). (3) A new evaluation metric AR-F is proposed.

d: BOUNDARY SENSITIVE NETWORK (BSN [21], 2018)

As we know, high-quality temporal action proposals should have several characteristics: a) Flexible temporal length. b) Precise temporal boundaries. c) Reliable confidence scores. But the existing methods cannot do well in these aspects at the same time. In order to solve the difficulties, T. Lin *et al.* [21] proposed BSN in 2018.

Briefly, BSN first locates the boundaries of the temporal action segments (the start node and the end node). And the boundaries nodes are directly combined into a temporal proposal. Then it extracts a 32-dimensional proposal-level feature based on the sequence of action confidence scores for each candidate proposal. Finally, based on the extracted feature of the proposal-level, it evaluates the confidence of the temporal proposals.

The main contributions are as follows: a) the novel framework can meet the requirements of the above 3 points at the same time. b) The modules of BSN are simple and flexible. The disadvantages are: a) the efficiency is not high enough because the process of feature extraction and confidence assessment is performed one by one for each temporal proposal. b) Insufficient semantic information. In order to ensure the efficiency of extracting action proposal feature, the 32-dimensional feature designed by BSN is relatively simple, but it also limits the confidence evaluation module to obtain more semantic information. c) This method has multi-stage. It doesn't optimize several parts of the network jointly.

e: BOUNDARY-MATCHING NETWORK (BMN [72], 2019)

In order to solve the shortcomings in BSN, the new temporal proposal confidence evaluation mechanism and boundary-matching mechanism were proposed also by T. Lin *et al.* [72] in 2019. BMN can generate the probability of one-dimensional boundary and the confidence map of two-dimensional BM simultaneously. Then it can evaluate the confidence scores of all possible temporal proposals densely.

The performance comparisons among the above temporal action proposal methods are shown in TABLE 1.

TABLE 1. Performance comparison: AR@AN = 200 on THUMOS14.

Methods	AR@200
S-CNN [14]	37.01% (C3D feature)
TAG [15]	39.61% (Two-stream feature)
TURN [16]	38.34% (C3D feature) 43.02% (Flow feature)
BSN+ Greedy-NMS [21]	43.61% (C3D feature) 52.23% (Two-stream feature)
BSN+ Soft-NMS [21]	45.55% (C3D feature) 53.21% (Two-stream feature)
BMN +Greedy NMS [72]	46.79% (C3D feature) 54.84% (Two-stream feature)
BMN+ Soft-NMS [72]	47.86% (C3D feature) 54.70% (Two-stream feature)

2) ONE-STAGE LOCALIZATION METHODS

The other type is one-stage framework which tackles proposal and classification simultaneously. For example, in 2017, T. Lin *et al.* proposed SSAD (single shot temporal action detection) [12] and the group of Li Feifei proposed SS-TAD (end-to-end, single-stream temporal action detection) [11]. Both of them are based on single-shot detector. Due to the similarity between temporal action localization and object detection, SSAD combines the characteristics of YOLO [73] and SSD [74] models of object detection. The general flows of SSAD are as follows. Using the pre-trained model, feature sequences are obtained which are the input of SSAD model. After processing, the model outputs the detection results. While SS-TAD improves training and testing performance using the semantic subtasks of temporal action localization as adjusted semantic constraints. So the effectiveness is better than SSAD. SS-TAD extracts feature using C3D, the same as SSAD. But SS-TAD adopts the anchor mechanism and the stacked GRU units. Recently, Fuchen Long *et al.* introduces GTAN (Gaussian Temporal Awareness Networks) [25] which integrates temporal structure to one-stage action localization. In GTAN, it introduces Gaussian kernels to optimize temporal scale of every action proposal dynamically.

In addition, some methods are based on sequential decision-making process that also belong to one-stage framework, such as [10], [13]. Literature [10] was the first one to propose an end to end approach to learning action detection in videos. In this article, it uses reinforcement learning to train an RNN-based agent. The agent can constantly observe the video frames and decide where to look next and when to generate an action prediction.

III. BENCHMARK DATASETS

Though there is not a standard benchmark for temporal action localization, most researchers use THUMOS14 [6] and ActivityNet [7]. Besides, there are several large-scale datasets for temporal action detection. For example, MEXaction2 [46], MutiTHUMOS [47], Charades [8] and AVA [9] and so on. The following paragraph mainly introduces several commonly used datasets.

A. THUMOS'14 [6]

THUMOS14 comes from THUMOS Challenge 2014. This dataset includes two tasks: action recognition and temporal action detection. Most papers are evaluated in this dataset. The THUMOS dataset has video-level annotations of 101action classes in its training, validation, and testing sets, and temporal annotations only for a subset of videos in the validation and testing sets for 20 classes.

The details for some fully-supervised learning methods are as follows: a) Training set: UCF101, 101 types of actions, a total of 13,320 trimmed video clips. b) Validation set: 1,010 untrimmed videos, 200 of them are labeled with temporal annotations. (3007 action segments, only 20 classes which can be used for the task of temporal action detection). c) Testing set: 1,574 untrimmed videos, 213 of them have temporal action annotations. (3,358 behavioral segments, only 20 classes which can be used for the task of temporal action detection.)

In a word, this dataset is challenging as some videos are relatively long (up to 26 minutes) and contain multiple action instances. The length of an action varies from less than a second to minutes significantly.

B. ActivityNet [7]

The ActivityNet dataset is the largest one which recently introduced benchmark for action recognition and action localization in untrimmed videos. This dataset only provides the link of YouTube video, but cannot download the videos directly. So we need to use the YouTube download tool in Python to download it automatically.

The ActivityNet1.3 consists 10,024 videos for training, 4,926 for validation, and 5044 for testing, with 200 activity classes, such as 'walking the dog', 'long jump', and 'vacuuming floor'. The total duration of the video is 648 hours. The ActivityNet 1.3 only contain 1.5 occurrences per video averagely and most videos simply contain single action category with 36% background on average.

This dataset contains a large number of natural videos that involve various human activities under a semantic taxonomy.

C. MEXaction2 [46]

MEXaction2 dataset contains two types of actions which are horse riding and bullfighting. This dataset consists of three parts: YouTube videos, horse riding videos in UCF101, and INA videos. Among them, the YouTube video clips and the horse riding video in UCF101 are trimmed short video clips that are used for the training set, while the INA videos are untrimmed long videos with a total length of 77 hours. The INA videos are divided into training, validation and test sets. There are 1,336, 310, and 329 action segments in the training set, validation set and testing set respectively.

In short, MEXaction2 dataset is characterized by the fact that the untrimmed videos are very long, and the annotations segments are only a small proportion of the total videos.

D. MUTITHUMOS [47]

MUTITHUMOS is a dense, multi-classes, frame wise labeled video dataset which includes 400 videos of 30 hours, 38,690 annotations of 65 classes. It has 1.5 labels per frame averagely, and 10.5 action classes of each video. It's an enhanced version of THUMOS. At present, we only saw the evaluation of this dataset in the paper "Learning Latent Super-Events to Detect Multiple Activities in Videos" in 2017.

E. CHARADES [8]

Charades is untrimmed videos which contain 9,848 indoor videos, (7985 training data, 1863 testing data), and 157 classes from 267 different people. Each video is about 30 seconds. Each video has multiple annotations and the start time and end time of each action.

F. AVA [9]

AVA is a spatio-temporally localized Atomic Visual Actions dataset. It contains 430 movie clips of 15 minutes' length which is annotated with 80 actions. There are 386,000 labeled segments, 614,000 labeled bounding boxes and 81,000 person tracks. There are totally 1.58M labelled actions and every person has multiple labels frequently.

Next we summarize and compare these datasets shown in TABLE 2.

IV. EVALUATION METRICS

A. BASIC CONCEPTS

In the problem of binary classification, TP represents True Positive, FP represents False Positive, TN represents True

TABLE 2. Comparison and summary of the datasets.

Dataset	Classes	Videos	Instances	Year
THUMOS14	101	15,000	3,000	2014
ActivityNet	200	20,000	7,600	2015
MEXaction2	2	-	1,975	2015
MUTITHUMOS	65	400	-	2017
Charades	157	9,848	-	2016
AVA	80	430	-	2018

TABLE 3. The logic detail of binary classification.

		Real results	
		1-true	**0-flase**
Predict results	**1-positive**	True Positive (TP)	False Positive (FP)
	0-negative	False Negative (FN)	True Negative (TN)

Negative, and FN represents False Negative. The four parameters are used to calculate many kinds of performance evaluation metrics. The logic details of four parameters are shown in TABLE 3.

In which, at the actual binary classification, the positive-1 label refers to the samples you are more concerned about, such as an action or an abnormal event.

1) ACCURACY

Accuracy is the proportion of classified samples correctly. It is used to evaluate the performance of the classifier.

$$accuracy = \frac{rP + TN}{TP + TN + FP + FN} = \frac{TP + TN}{ALL} \tag{1}$$

2) RECALL

Recall is the coverage of predicting correctly. Specifically, recall is that how many real positive samples in the testing set were identified. The formula is as follows.

$$recall = \frac{TP}{TP + FN} \tag{2}$$

3) PRECISION

Specifically, precision is the percentage of the predicted real positive samples in predicted results. The formula is as follows.

$$Precision = \frac{TP}{TP + FP} = \frac{TP}{n} \tag{3}$$

In which, n is the sum of True Positive and False Positive, and n is also the total number of samples identified by the system.

4) INTERSECTION-OVER-UNION (IoU)

IoU can be understand as the overlap between the predicted detection box by the model and the ground truth for the object detection in images. In fact, it is the accuracy of detection. The calculation formula is the intersection of Detection Result and Ground Truth compared to their union.

$$IoU = \frac{predicted\ detection\ box \cap ground\ truth}{predicted\ detection\ box \cup ground\ truth} \tag{4}$$

IoU is used to check whether the IoU between the predicted result and the ground truth is greater than a predicted threshold. We often set 0.5 as the threshold. If the IoU is greater than 0.5, the object will be identified as "detected successfully", otherwise it will be identified as "missed". In temporal action detection, IoU is changed into t-IoU for time which has only one dimension.

B. EVALUATION METRICS

C. AVERAGE RECALL (AR)

AR is the evaluation metric for temporal action proposals generation. Because temporal action proposal generation doesn't need to classify, it only needs to find the proposal. Therefore, whether the temporal proposals we find are complete can be used to evaluate the performance of the method. So we often use AR for the judgment.

$$AR = \frac{sum\ of\ the\ videos\ recalled}{total\ number\ of\ videos} \tag{5}$$

D. MEAN AVERAGE PRECISION (mAP)

In the task of temporal action localization, mAP is the evaluation metric which we most commonly used. In general, we compare mAPs in the case of t-IoU = 0.5.

To say simply, Precision (P) is the degree of correct detection in a single class of a given video. For example, for a given single video, precision in Class C is shown in the formula.

$$P = \frac{TP}{TP + FP} = \frac{number\ of\ predicted\ correct\ proposals}{total\ number\ of\ predicted\ proposals} \tag{6}$$

Because there are many videos in the testing set, Average Precision (AP) is the average precision of all videos in Class C. At the same time, because there are also many classes corresponding to the testing set videos, so Mean Average Precision is the average precision of all classes in all testing videos.

$$mAP = \frac{the\ sum\ of\ average\ precision\ of\ all\ classes}{total\ number\ of\ videos\ in\ testing\ set} \tag{7}$$

In a word, under a certain t-IoU, P is the accuracy of predicted proposals of a certain Class C in a video. AP is the average accuracy of the predicted proposals of all classes in a video. MAP is the mean of the average accuracy of the predicted proposals of all classes in all testing videos. Following the standard evaluation protocol, almost all papers report mAP at different thresholds of t-IoU.

V. RECENT METHODS AND DEVELOPMENTS

A. FULLY-SUPETVISED TEMPORAL ACTION LOCALIZATION (F-TAL)

1) FULLY-SUPETVISED LEARNING

Fully-supervised learning is a process to train an intelligent algorithm to map the input data into labels, where each training data has its corresponding label indicating its ground truth. Classification and regression which we often study are the representatives of supervised learning. In the task of temporal action localization, full supervision employs the labels of training set that contain both the video-level category labels and the temporal annotations information of the action segment (including the start and the end time of the action).

2) CURRENT REPRESENTATIVE METHODS

Many of the methods (such as S-CNN [14] and PSDF [18]) generate proposals by sliding window, and classify them into C + 1 classes that is C action classes and one background class. Among them, S-CNN uses a multi-stage CNN for temporal action localization to capture the robust video feature representation. For precise boundaries, the CDC [19] (Convolutional-De-Convolutional) network and the TPC-Net [20] (Temporal Preservation Convolutional network) are proposed for frame-level action predictions. BSN [21] (Boundary Sensitive Network) is recently proposed to locate temporal boundaries which are further integrated into action proposals. Next year, the authors of BSN also propose a new temporal proposal confidence evaluation mechanism and boundary-matching mechanism BMN [72]. BMN can generate the probability of one-dimensional boundary and the confidence map of two-dimensional BM simultaneously. For the completeness of the proposals, SSN [22] introduces structured temporal pyramids with decoupled classifiers for classifying actions and determining completeness. Furthermore, some region-based methods (such as R-C3D [23] and TAL-Net [24]) propose to generalize the methods for 2D object detection to 1D temporal action localization. Recently, TSA-Net [26] and Gaussian temporal modeling [25] are proposed for accurate action localization. The following are the performance comparisons. For simplicity and fairness, we take performance comparison of mAP@tIoU = 0.5 on the THUMOS14 dataset and publications of various representative methods, as shown in the TABLE 4.

In recent years, with the introduction of various new networks, the accuracy has reached the latest 46.9% [26]. Of course, there is still a certain gap with object detection in images, which is why it is difficult to commercialize on a large scale currently. But we can believe that with the continuous progress of technology, the accuracy will achieve a breakthrough.

TABLE 4. comparisons: mAP@tIoU = 0.5 on the THUMOS14.

	Fully-supervised Methods	mAP@tIoU= 0.5	Published
Two-stage frame work	S-CNN [14]	19.0%	CVPR 2016 (the highest)
	PSDF [18]	18.8%	CVPR 2016
	CDC [19]	23.3%	CVPR 2017
	TPN [20]	27.6%	AAAI 2017
	TAG [15]	28.2%	CVPR 2017
	TURN [16]	25.6%	CVPR 2017
	SSN [22]	29.8%	ICCV 2017
	CBR [27]	31.0%	BMVC 2017 (the highest)
	R-C3D [23]	28.9%	ICCV 2017
	ETP [28]	34.2%	ICMR 2018
	TAL-Net [24]	42.8%	CVPR 2018 (the highest)
	BSN+SCNN [21] BSN+UntrimmedNet [21]	29.4% 36.9%	ECCV 2018
	BMN+SCNN [72] BMN+UntrimmedNet [72]	32.2% 38.8%	arXiv 2019
	TSA-Net [26]	46.9%	arXiv 2019
One-stage frame work	End-to-End learning [10]	17.1%	CVPR 2016
	SMS [75]	17.8%	CVPR 2017
	SSAD (C3D) [12]	24.6%	ACM MM 2017
	SS-TAD (C3D) [11]	29.2%	BMVC 2017
	GTAN (C3D)[25] GTAN (P3D)[25]	37.9% 38.8%	CVPR 2019

B. WEAKLY-SUPETVISED TEMPORAL ACTION LOCALIZATION (W-TAL)

According to the above section, we know that current fully-supervised learning techniques have achieved great success in temporal action localization. Because many existing techniques rely on trimmed videos as their inputs, such as UCF101, and they have these precise temporal annotations. But in the realistic scenario, most of the videos are untrimmed and contain many frames that are not relevant to target actions. So it is very difficult to require the temporal annotations. The specific reasons are summarized as follows: a) frame-level annotations for every action instance is expensive and time-consuming. b) There is no sensible definition about the temporal action exactly, so these temporal annotations may be subjective by different people.

So weakly-supervised learning methods have been more and more popular.

1) WEAKLY-SUPETVISED LEARNING

Let's take a look at weakly- supervised learning. There are three types of weakly-supervised learning [40]. a) Incomplete supervised, that is only a little subset of training data are labeled, while the other data have no labels. For example, in image classification, we can easily get a large number of images from the internet, but only a small number of images have annotations due to the expensive cost by human. b) Inexact supervised, that is the training data only have coarse-grained labels. We also take images classification as an example. We usually have image-level labels but not object-level labels. c) Inaccurate supervised, that is the labels given to us are not always the ground-truth. For example, when the image annotator is tired or careless, or some images are very difficult to classify, such situation will happen.

According to the above, since weakly-supervised temporal time localization only has video-level labels but not frame-level temporal annotations in the training process, it belongs to the second type of weak supervision which is inexact supervised.

2) CURRENT REPRESENTATIVE METHODS

There are only a few methods based on weakly-supervised which only rely on video-level class labels to temporal action localization. Motivated by the weakly-supervised object detection in images, UntrimmedNet [29] and Hide-and-seek [30] are studied by researchers. UntrimmedNet is the first one to propose action recognition and action detection with weak supervision. It's an end-to-end model to learn single label action classification and detection. STPN [31] is a deep neural network based on classification. The general structure of the network is as follows: the video is divided into N segments, and the attention module can identify a sparse subset of the key segments. Then we can obtain the importance of each segment in the process of predicting classification label. Thereby, it can generate the corresponding category labels and interval suggestions by adaptive temporal pooling. AutoLoc [32] attempts to predict the temporal boundary directly that is different from the previous weakly-supervised temporal action detection according to the threshold on the CAS. The main idea is that the average score outside the action segment is encouraged to be lower than the inside for score of action category. W-TALC [33] introduces a novel function to get K-max Multiple Instance Learning and unearth co-activity relationship between the localized instances of the same class. To solve the problem of fragmentation of video frames that the classifier cares about and the completeness of the action, Hide-and-seek [30] randomly hides some frames to force residual attention to learn the relatively low discrimination video frames in each training. Although, it doesn't guarantee the discovery of new parts at each training. The result shows that this method works well for spatial object detection but is not good for temporal action localization. Step-by-step erasion, one-by-one collection [34] erases and trains multiple classifiers step by step, and merges the prediction segments of each classifier directly. The performance is better, but it costs more time and computation. Afterwards, CMCS [35] proposes a multi-branch network architecture with diversity loss for action completeness modeling. At the same time, they propose a scheme generating a hard negative video for separating contexts. Although the main point of this article is not the background class, it inspires the next subsequent three works that are BaSNet [36], background modeling [37], and LPAT [38]. Without considering the background category, the background frames were misclassified into action categories, resulting in a large number of FPs. In BaSNet, in order to construct negative samples of the background class, an attention module was introduced in another network to suppress the background response. Two other works take into account background class from different aspects, and suppress the influence of background effectively. Finally, they all improve the accuracy of the localization.

TABLE 5. Performance comparison: mAP@tIoU = 0.5 on the THUMOS14.

Weakly-supervised Methods	mAP@tIoU=0.5	Published
Domain transfer from web images	4.4%	ACM MM2015
UntrimmedNet [29]	13.7%	CVPR 2017
Hide-and-Seek [30]	6.84%	ICCV 2017
STPN [31]	16.2% (UntrimmedNet feature) 16.9% (I3D feature))	CVPR 2018
Step-by-step erasion [34]	15.9%	ACM MM2018
AutoLoc [32]	21.2%	ECCV 2018
W-TALC [33]	18.8% (UntrimmedNet feature) 22.8%（I3D feature）	ECCV 2018
CPMNet	16.1%	ACCV 2018
STAR [76]	23.0%	AAAI 2019
CMCS [35]	19.9% （UntrimmedNet feature) 23.1%(I3D feature)	CVPR 2019
LPAT [38]	22.6% (Untrimmed feature) 27.9% （I3D feature）	Arxiv 2019
Background Modeling [37]	26.8%	ICCV2019
BaSNet [36]	25.1% (UntrimmedNet feature) 27.0% (I3D feature)	AAAI 2020

The following are the performance comparisons on the THUMOS14 dataset, the same standard as fully-supervised learning and publications of various representative methods, as shown in the TABLE 5.

3) INSIGHTS ON THE PROBLEM OF W-TAL

Recently, multiple instance learning (MIL) has been used for W-TAL. Instead of learning with s set of instances that are individually labeled, a MIL model receives a set of labeled bags, each of them containing many instances. If we consider action instances in a video as a bag, and video-level annotation as the label, then the W-TAL can be formulated as a process of multiple instance learning.

Temporal class activation mapping (T-CAM) or Class activation sequence (CAS) is another recently group of methods for W-TAL. The CNN visualization has shown that the convolution layer of a CNN performs as the action detectors, although there is no supervision on the location of the activities. Class activation sequence elucidates that a CNN enables to have localization ability despite being trained on video-level labels. In addition, some other research on W-TAL were inspired by weakly supervised object detection, such as interactive annotation and generative adversarial training.

All in all, weakly-supervised learning reduces the costs of labor and time, but also increases the difficulty of temporal detection. But for most video clips in most action categories, the results seem to be good. Of course, there is still plenty of room for improvements.

VI. FUTURE DIRECTIONS AND TRENDS

The application of temporal action localization will be more and more wide actually, and the future trends may focus on but is not limited to the followings.

a) Precision and efficiency improvements. Compared the method two-stream and 3D convolution, two-stream is more accurate but less efficient than the latter. How to take the advantages of both of them better is a possible research direction in

b) Action detection will extend from temporal action detection to spatio-temporal action detection [39]. That is to say, we should detect from one-dimensional temporal interval to two-dimensional spatio-temporal box that can detect actions more comprehensively.

c) Action detection of videos online. That is a process of dealing with a video stream which needs to detect the category of action online, but cannot know the content after the detection time. The setting of online is more conform to the requirements of surveillance videos which need real-time detection or early warning, such as anomaly detection [41].

d) Weakly-supervised learning of temporal action localization will become more and more popular. In many tasks, it's difficult to obtain full supervision information due to the high cost of the data labeling process.

e) Video is a multi-modal data which contains image and audio. Whether to use audio information to assist temporal action localization is direction worth considering. As Aytar *et al.* had used image-assisted audio analysis [45].

VII. CONCLUSION

In this article, we conduct a comprehensive overview of temporal action localization. We analyze related techniques from time division: traditional methods and deep learning methods. Next we summarize the benchmark datasets and analyze the evaluation metrics. Then we review the recent developments of temporal action localization from fully-supervised learning to weakly-supervised learning methods. Anyway, we have tried to give the relevance and current situation of temporal action localization. At the same time, we hope to help some readers who are interested in temporal action localization.

Temporal action localization, as a hot topic in video understanding, is very complicated and challenging. However, in the near future, we believe that the results can be improved and the task will become easier with the exploiting of deep learning techniques.

REFERENCES

[1] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013.

[2] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1725–1732.

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

[4] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4489–4497.

[5] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6299–6308.

[6] Y.-G. Jiang, J. Liu, A. R. Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar, "THUMOS challenge: Action recognition with a large number of classes," in *Proc. Int. Workshop Competition Action Recognit. Large Number Classes (ECCV)*, 2014. [Online]. Available: http://crcv.ucf.edu/THUMOS14/

[7] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, "ActivityNet: A large-scale video benchmark for human activity understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 961–970.

[8] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," in *Proc. ECCV*, 2016, pp. 510–526.

[9] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar, C. Schmid, and J. Malik, "AVA: A video dataset of spatio-temporally localized atomic visual actions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6047–6056.

[10] S. Yeung, O. Russakovsky, G. Mori, and L. Fei-Fei, "End-to-End learning of action detection from frame glimpses in videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2678–2687.

[11] S. Buch, V. Escorcia, B. Ghanem, and J. C. Niebles, "End-to-end, single-stream temporal action detection in untrimmed videos," in *Proc. Brit. Mach. Vis. Conf.*, 2017, p. 7.

[12] T. Lin, X. Zhao, and Z. Shou, "Single shot temporal action detection," in *Proc. ACM Multimedia Conf. (MM)*. New York, NY, USA: ACM, 2017, pp. 988–996.

[13] H. Alwassel, F. Caba Heilbron, and B. Ghanem, "Action search: Spotting actions in videos and its application to temporal action localization," in *Proc. ECCV*, Sep. 2018, pp. 253–269.

[14] Z. Shou, D. Wang, and S.-F. Chang, "Temporal action localization in untrimmed videos via multi-stage CNNs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1049–1058.

[15] Y. Xiong, Y. Zhao, L. Wang, D. Lin, and X. Tang, "A pursuit of temporal accuracy in general activity detection," 2017, *arXiv:1703.02716*. [Online]. Available: http://arxiv.org/abs/1703.02716

[16] J. Gao, Z. Yang, C. Sun, K. Chen, and R. Nevatia, "TURN TAP: Temporal unit regression network for temporal action proposals," 2017, *arXiv:1703.06189*. [Online]. Available: http://arxiv.org/abs/1703.06189

[17] S. Buch, V. Escorcia, C. Shen, B. Ghanem, and J. C. Niebles, "SST: Single-stream temporal action proposals," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2911–2920.

[18] J. Yuan, B. Ni, X. Yang, and A. A. Kassim, "Temporal action localization with pyramid of score distribution features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3093–3102.

[19] Z. Shou, J. Chan, A. Zareian, K. Miyazawa, and S.-F. Chang, "CDC: Convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5734–5743.

[20] K. Yang, P. Qiao, D. Li, S. Lv, and Y. Dou, "Exploring temporal preservation networks for precise temporal action localization," in *Proc. AAAI*, 2018, pp. 7477–7484.

[21] T. Lin, X. Zhao, H. Su, C. Wang, and M. Yang, "BSN: Boundary sensitive network for temporal action proposal generation," in *Proc. ECCV*, 2018, pp. 3–21.

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

[23] H. Xu, A. Das, and K. Saenko, "R-C3D: Region convolutional 3D network for temporal activity detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5783–5792.

[24] Y.-W. Chao, S. Vijayanarasimhan, B. Seybold, D. A. Ross, J. Deng, and R. Sukthankar, "Rethinking the faster R-CNN architecture for temporal action localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1130–1139.

[25] F. Long, T. Yao, Z. Qiu, X. Tian, J. Luo, and T. Mei, "Gaussian temporal awareness networks for action localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 344–353.

[26] G. Gong, L. Zheng, K. Bai, and Y. Mu, "Scale matters: Temporal scale aggregation network for precise action localization in untrimmed videos," 2019, *arXiv:1908.00707*. [Online]. Available: http://arxiv.org/abs/1908.00707

[27] J. Gao, Z. Yang, and R. Nevatia, "Cascaded boundary regression for temporal action detection," in *Proc. Brit. Mach. Vis. Conf.*, 2017, pp. 1–11.

[28] H. Qiu, Y. Zheng, H. Ye, Y. Lu, F. Wang, and L. He, "Precise temporal action localization by evolving temporal proposals," in *Proc. ACM Int. Conf. Multimedia Retr.*, Jun. 2018, pp. 388–396.

[29] L. Wang, Y. Xiong, D. Lin, and L. Van Gool, "UntrimmedNets for weakly supervised action recognition and detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4325–4334.

[30] K. K. Singh and Y. J. Lee, "Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3544–3553.

[31] P. Nguyen, B. Han, T. Liu, and G. Prasad, "Weakly supervised action localization by sparse temporal pooling network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6752–6761.

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

[33] S. Paul, S. Roy, and A. K. Roy-Chowdhury, "W-TALC: Weakly-supervised temporal activity localization and classification," in *Proc. ECCV*, 2018, pp. 588–607.

[34] J.-X. Zhong, N. Li, W. Kong, T. Zhang, T. H. Li, and G. Li, "Step-by-step erasion, One-by-one collection: A weakly supervised temporal action detector," in *Proc. ACM Multimedia Conf. Multimedia Conf. (MM)*, 2018, pp. 35–44.

[35] D. Liu, T. Jiang, and Y. Wang, "Completeness modeling and context separation for weakly supervised temporal action localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1298–1307.

[36] P. Lee, Y. Uh, and H. Byun, "Background suppression network for weakly-supervised temporal action localization," in *Proc. AAAI*, 2020, pp. 1–8.

[37] P. Nguyen, D. Ramanan, and C. Fowlkes, "Weakly-supervised action localization with background modeling," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5502–5511.

[38] X. Lin, Z. Shou, and S.-F. Chang, "LPAT: Learning to predict adaptive threshold for weakly-supervised temporal action localization," 2019, *arXiv:1910.11285*. [Online]. Available: https://arxiv.org/abs/1910.11285

[39] L. Song, S. Zhang, G. Yu, and H. Sun, "TACNet: Transition-aware context network for spatio-temporal action detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11987–11995.

[40] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *Nat. Sci. Rev.*, vol. 5, no. 4, pp. 44–53, 2018.

[41] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018.

[42] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human action classes from videos in the wild," Univ. Central Florida, Orlando, FL, USA, Tech. Rep. CRCV-TR-12-01, 2012.

[43] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2556–2563.

[44] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The kinetics human action video dataset," 2017, *arXiv:1705.06950*. [Online]. Available: http://arxiv.org/abs/1705.06950

[45] Y. Aytar, "SoundNet: Learning sound representations from unlabeled video," in *Proc. NIPS*, 2016, pp. 892–900.

[46] C. Michel and B. P. Jenny. *MEXaction2: Action Detection and Localization Dataset*. Accessed: 2015. [Online]. Available: http://mexculture.cnam.fr/Datasets/mex+action+dataset.html

[47] A. Piergiovanni and M. S. Ryoo, "Learning latent super-events to detect multiple activities in videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5304–5313.

[48] C. Schmid and S.-F. Chang, "ActivityNet large scale activity recognition challenge," in *Proc. CVPR*, 2017, pp. 961–970. [Online]. Available: http://activity-net.org/challenges/2017/index.html

[49] S. Vishwakarma and A. Agrawal, "A survey on activity recognition and behavior understanding in video surveillance," *Vis. Comput.*, vol. 29, no. 10, pp. 983–1009, 2013.

[50] Y. Gao, S. S. Vedula, C. E. Reiley, N. Ahmidi, B. Varadarajan, H. C. Lin, L. Tao, L. Zappella, B. Bejar, D. D. Yuh, C. C. G. Chen, R. Vidal, S. Khudanpur, and G. D. Hager, "JHU-ISI gesture and skill assessment working set (JIGSAWS): A surgical activity dataset for human motion modeling," *MICCAI Workshop (M2CAI)*, vol. 3, 2014, p. 3.

[51] Y. Jae Lee, J. Ghosh, and K. Grauman, "Discovering important people and objects for egocentric video summarization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1346–1353.

[52] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[53] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. CVPR*, 2015, pp. 2625–2634.

[54] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4694–4702.

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

[56] K. Schindler and L. van Gool, "Action snippets: How many frames does human action recognition require?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[57] S. Satkin and M. Hebert, "Modeling the temporal extent of actions," in *Proc. ECCV*, 2010, pp. 536–548.

[58] B. Fernando, C. Tan, and H. Bilen, "Weakly supervised Gaussian networks for action detection," in *Proc. CVPR*, 2019, pp. 537–546.

[59] V. Escorcia, F. Caba Heilbron, J. C. Niebles, and B. Ghanem, "DAPs: Deep action proposals for action understanding," in *Proc. ECCV*, 2016, pp. 768–784.

[60] L. Wang, Y. Qiao, X. Tang, and L. V. Gool, "Actionness estimation using hybrid fully convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2708–2717.

[61] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. NIPS*, 2012, pp. 1097–1105.

[62] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, vol. 2, Sep. 1999, pp. 1150–1157.

[63] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[64] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 886–893.

[65] J. C. Niebles, H. Wang, and F. Li, "Unsupervised learning of human action categories using spatial-temporal words," *Int. J. Computer. Vis.*, vol. 79, pp. 299–318, Sep. 2008.

[66] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proc. IEEE Int. Workshop Vis. Surveill. Perform. Eval. Tracking Surveill.*, Beijing, China, Oct. 2005, pp. 65–72.

[67] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *Int. J. Comput. Vis.*, vol. 103, no. 1, pp. 60–79, May 2013.

[68] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[69] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real- time object detection with region proposal networks," in *Proc. Adv. NIPS*, 2015, pp. 91–99.

[70] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[71] I. Rodríguez-Moreno, J. M. Martinez-Otzeta, B. Sierra, I. Rodriguez and E. Jauregi, "Video activity recognition: State-of-the-art," *Sensors*, vol. 19, p. 3160, Jan. 2019.

[72] T. Lin, X. Liu, X. Li, E. Ding, and S. Wen, "BMN: Boundary-matching network for temporal action proposal generation," 2019, *arXiv:1907.09702*. [Online]. Available: http://arxiv.org/abs/1907.09702

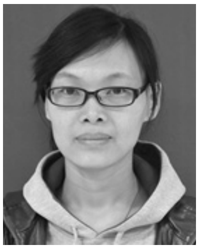
[73] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[74] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot MultiBox detector," in *Proc. ECCV*, 2016, pp. 21–37.

[75] Z. Yuan, J. C. Stroud, T. Lu, and J. Deng, "Temporal action localization by structured maximal sums," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3684–3692.

[76] Y. Xu, C. Zhang, Z. Cheng, J. Xie, Y. Niu, S. Pu, and F. Wu, "Segregated temporal assembly recurrent networks for weakly supervised multiple action detection," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, pp. 9070–9078, Jul. 2019.

[77] Z. Zou, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," 2019, *arXiv:1905.05055*. [Online]. Available: http://arxiv.org/abs/1905.05055



HUIFEN XIA was born in Nantong, Jiangsu, China. She received the B.S. degree in information and computational science from Jiangsu Ocean University, in 2006, and the M.S. degree in system engineering from Jiangsu University, China, in 2008, where she is currently pursuing the Ph.D. degree with the Department of Computer Science and Communication Engineering. She has high interests in machine learning and deep learning. Her research interests include computer vision and video understanding, especially video temporal action localization.

YONGZHAO ZHAN was born in Sanming, Fujian, China, in 1962. He received the B.S. degree from Fuzhou University, China, in 1984, the M.S. degree from Jiangsu University, China, in 1990, and the Ph.D. degree from Nanjing University, China, in 2000, all in computer science. He is currently a Professor with the School of Computer Science and Communication Engineering, Jiangsu University. He has authored over 80 articles. His research interests include big data, multimedia, and the Internet of Vehicles. He was a recipient of the Science and Technology Progress Award from the Government of Zhenjiang, in 2006, and from the Government of Jiangsu, in 2013.