electronics

MDPI

Article

Advancing Temporal Action Localization with a Boundary Awareness Network

Jialiang Gu *, Yang Yi and Min Wang

School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510006, China
* Correspondence: gujliang@mail2.sysu.edu.cn

Abstract: Temporal action localization (TAL) is crucial in video analysis, yet presents notable challenges. This process focuses on the precise identification and categorization of action instances within lengthy, raw videos. A key difficulty in TAL lies in determining the exact start and end points of actions, owing to the often unclear boundaries of these actions in real-world footage. Existing methods tend to take insufficient account of changes in action boundary features. To tackle these issues, we propose a boundary awareness network (BAN) for TAL. Specifically, the BAN mainly consists of a feature encoding network, coarse pyramidal detection to obtain preliminary proposals and action categories, and fine-grained detection with a Gaussian boundary module (GBM) to get more valuable boundary information. The GBM contains a novel Gaussian boundary pooling, which serves to aggregate the relevant features of the action boundaries and to capture discriminative boundary and actionness features. Furthermore, we introduce a novel approach named Boundary Differentiated Learning (BDL) to ensure our model's capability in accurately identifying action boundaries across diverse proposals. Comprehensive experiments on both the THUMOS14 and ActivityNet v1.3 datasets, where our BAN model achieved an increase in mean Average Precision (mAP) by 1.6% and 0.2%, respectively, over existing state-of-the-art methods, illustrate that our approach not only improves upon the current state of the art but also achieves outstanding performance.

Keywords: action boundary detection; Gaussian boundary module; video understanding; temporal action localization

check for updates

Citation: Gu, J.; Yi, Y.; Wang, M. Advancing Temporal Action Localization with a Boundary Awareness Network. *Electronics* **2024**, *13*, 1099. https://doi.org/10.3390/electronics13061099

Academic Editor: Daniel Morris

Received: 23 February 2024
Revised: 14 March 2024
Accepted: 14 March 2024
Published: 16 March 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

With multiple videos being submitted to social media every day, an effective and efficient video understanding algorithm is becoming increasingly important. Temporal action localization (TAL) plays a crucial role in video interpretation, particularly for applications such as video surveillance, content analysis and summarization, and human–computer interaction. In long, untrimmed videos, TAL is essential to not only detect temporal proposal boundaries of action instances but also to identify action categories.

TAL has great similarity with object detection; TAL is aimed at accurately generating temporal proposals to segment action instances, while object detection is aimed at precisely producing spatial proposals to delineate object instances. Therefore, many methods of TAL are based on the extension of the object detection framework, including anchor-based [1–6], actionness-guided [7–11], and the anchor-free (boundary-based) method [12,13], which is inspired by the flexibility and efficiency of anchor-free object detectors, such as CornerNet [14], FCOS [15], and BorderDet [16]. Presently, anchor-based methods in temporal action localization (TAL) are generally classified into two distinct types: two-stage framework [2–4] with higher accuracy and one-stage framework [1,5,6] with higher efficiency. The two-stage framework studies proposal and classification. The former is divided into the task of temporary action proposal (TAP), and the latter is more similar to action recognition.

Methods that are anchor-free and known for their accuracy in defining boundaries, their adaptability in duration, and produce a single proposal at each temporal point

throughout the video sequence. This proposal is represented by a pair of values indicating the distance from the current location to the start and end times. Among these, the Anchor-Free Saliency-based Detector (AFSD) has made notable strides, pioneering the use of frame-level features and saliency cues to enhance boundary detection. Although the above methods have achieved good results, there is still some distance from the goal of accurately localizing the action, the most difficult aspect of which is the ambiguity of action boundaries. Due to the continuity of the action, the action does not start suddenly at a certain moment, but "gradually". If the "action value" is used to distinguish the beginning (or end) of an action, then the "action value" accumulates until the threshold (i.e., the moment of the action) before the action begins (or after it ends). Therefore, the initiation point of actions (which is identical to the termination time) must exhibit two critical characteristics: (1) Relatively small variation in the characteristics of the boundary area. (2) There is an objective threshold, which is used to distinguish the boundaries of actions.

To this end, we introduce the boundary awareness network (BAN) as a refinement to the existing anchor-free framework for temporal action localization (TAL), focusing on achieving a more accurate detection of action boundaries within videos. This refinement is executed through a comprehensive architecture comprising a feature encoding network, a coarse pyramidal detection module, and a fine-grained detection mechanism powered by the Gaussian boundary module (GBM). Compared to the anchor-free method AFSD, the BAN introduces innovative approaches for boundary detection, aiming to address the ambiguities of action boundaries more effectively; while the AFSD laid the groundwork for anchor-free TAL by simplifying the detection process and reducing dependency on hyper-parameters, the BAN builds upon this by enhancing the precision of boundary detection through the GBM and a novel learning strategy, Boundary Differentiated Learning (BDL).

Feature encoding network. Utilizing the I3D [17], renowned for its effectiveness in capturing rich, dynamic features from video data, this network serves as the cornerstone of our approach. By extracting robust spatial–temporal features, it lays a solid foundation for the precise detection and localization of actions within videos.

Coarse pyramidal detection module. To address the spatial–temporal dimensions of video data comprehensively, we employ a feature pyramid network equipped with various temporal convolutions. This module is pivotal in generating initial action proposals by amalgamating information across multiple layers, thereby facilitating a multi-scale understanding of potential action instances.

Fine-grained detection. Then, rough estimates of the start and end boundary distances are derived through projection. After that, the predicted temporal regions are used to generate fine-grained features with the Gaussian boundary module, which utilizes Gaussian boundary pooling to precisely locate the action boundary. Since the action is continuous, its boundary has a fuzzy property. Accurate action boundary detection requires discriminative boundary features. However, the core of obtaining such boundary features is how to correlate the moment boundary features with relevant features. A useful but non-differentiable method is to extract the moment-level boundary feature from the frame feature with the maximum activation value; while the boundary feature contains the most discriminating features, most useful information is lost. For example, the features around the salient features also contain boundary information.

Gaussian Boundary Module. To tackle the challenges in capturing high-quality boundary features for action prediction, we introduce the Gaussian boundary module (GBM). This module features a novel boundary pooling method named Gaussian boundary pooling (GBP), employing a Gaussian kernel function to aggregate features around the action boundary effectively. The kernel, centered on the boundary time stamp from coarse predictions, ensures that features closer to the boundary have a higher influence, optimizing the quality of boundary features extracted. It is important to note the distinction between our Gaussian kernels and those used in GTAN, which aims to represent action proposals of varying durations, whereas our focus is on extracting moment range boundary features for precise action localization.

Boundary Differentiated Learning. To further enhance our model's ability to discern complex actions from background clutter in videos, we propose a novel strategy named Boundary Differentiated Learning (BDL). This strategy innovatively transforms the classification task of distinguishing action boundaries into a regression task focused on identifying the 'actionability' of a scene. By re-scaling the ground truth labels to guide the model, BDL allows for a more nuanced understanding of action presence, facilitating the discrimination between action and non-action segments. Our extensive experiments on datasets such as THUMOS14 and ActivityNet-1.3 demonstrate the effectiveness of the BDL strategy, significantly improving our model's performance and establishing new state-of-the-art results.

Our key contributions can be outlined as follows:

1. In this study, our enhanced anchor-free framework for temporal action localization modestly increases computational demands relative to the AFSD method, yet it offers a more precise detection of action boundaries. This methodical refinement bolsters the framework's precision without significantly impacting computational efficiency, thereby advancing the state of the art in temporal action localization.

2. For a more accurate delineation of action boundaries, we have developed a unique boundary feature extraction method using the Gaussian boundary module. This is utilized in conjunction with coarse pyramidal detection to yield more detailed predictions. Additionally, we have devised a Boundary Contrastive Learning approach to further refine the accuracy of boundary information.

3. In our thorough assessments on the THUMOS14 and ActivityNet v1.3 datasets, the boundary awareness network (BAN) has shown advancements in mean Average Precision (mAP) when compared with the existing state-of-the-art methods, underscoring the enhanced performance of our proposed approach.

2. Related Work

Anchor-based. Current anchor-based approaches in the field can be categorized into two distinct frameworks: the one-stage and the two-stage framework. The two-stage framework includes TURN-TAP [2], R-C3D [3], TAL-Net [4], etc. TURN-TAP [2] cuts the untrimmed video into multiple non-overlapping clips, uses the pyramid method to construct unit features, and independently regresses the temporal boundary on each unit. R-C3D [3], taking inspiration from Faster-RCNN [18], proposes an end-to-end network to generate TAP and action classification in two stages. TAL-Net [4] proposes multi-scale anchors to adapt to the diversified duration of action instances. The main drawback of the two-stage framework is that the boundaries of action instances have been fixed during the classification step. To address this issue, a one-step framework has been proposed in recent years, including SSAD [1], GTAN [5], and PBR-Net [6]. SSAD [1] uses 1D temporal convolution to generate multiple temporal action anchors, and captures actions with different durations from three hierarchical levels. GTAN [5] represents actions with different durations through the mixture of Gaussian kernels. PBR-Net [6] designs three series regression modules to modify the boundary information gradually and integrates the fine-grained frame-level features method and anchor-based method.

The above anchor-based methods heavily rely on predefined anchors, which lack flexibility and are sensitive to anchor parameters. To address these problems, A2Net [19] combines anchor-based and anchor-free method to accurately detects TAP. However, it depends on predefined anchors to some extent. Our method differs, in that it does not require predefined anchors, and is thus more efficient and flexible.

Actionness-guided. Anchor-based TAL method can be regarded as a top–down algorithm, whereas the actionness-guided TAL method is bottom–up. It indicates the probability of potential actions utilizing the frame-level actionness score, then combines all start and end moments into proposals, and, finally, performs category predictions for each proposal. Compared with the anchor-based method, the actionness-guided method detects proposals in a more flexible manner. SSN [7] models the different stages (start,

course, and end) of the action, and evaluates the completeness of proposals. CDC [11] predicts the category of action frame by frame, which makes the boundary of action more accurate. BSN [8] locates action boundaries via actionness score and starting and ending probability. BMN [9], an extension of BSN, combines the temporary boundary probability sequence and bounding-matching mechanism to get the final proposals, and designs the threshold to screen proposals. However, the boundary matching strategy of BMN only uses the low-level boundary information and ignores the overall information of action, which performs poorly in dealing with complex actions and fuzzy backgrounds. To overcome this limitation, DBG [10] assesses comprehensive boundary confidence maps across all proposals and introduces features sensitive to action for a thorough evaluation of action completeness.

In contrast to these approaches, anchor-free methods like SALAD [13] and the AFSD [12] demonstrate efficiency and flexibility by eschewing predefined anchors. The AFSD, in particular, has pioneered an entirely anchor-free framework for TAL, emphasizing moment-level feature identification for action boundaries. Despite its innovations, the AFSD encounters challenges in capturing comprehensive boundary features, often losing valuable information surrounding salient features.

Building on the AFSD's foundation, we introduce the boundary awareness network (BAN) equipped with the Gaussian boundary module (GBM) to enhance action boundary localization precision. Unlike the AFSD, the BAN leverages Gaussian boundary pooling (GBP) within the GBM to aggregate features around action boundaries more effectively, ensuring that high-quality boundary features are captured without losing peripheral information. This targeted approach allows the BAN to address the limitations observed in both anchor-based and actionness-guided methods, marking a significant step forward in the domain of anchor-free temporal action localization.

3. Method

In this section, the detailed structure of our newly developed boundary awareness network (BAN) is discussed. The comprehensive architecture tailored for temporal action localization is depicted in Figure 1.

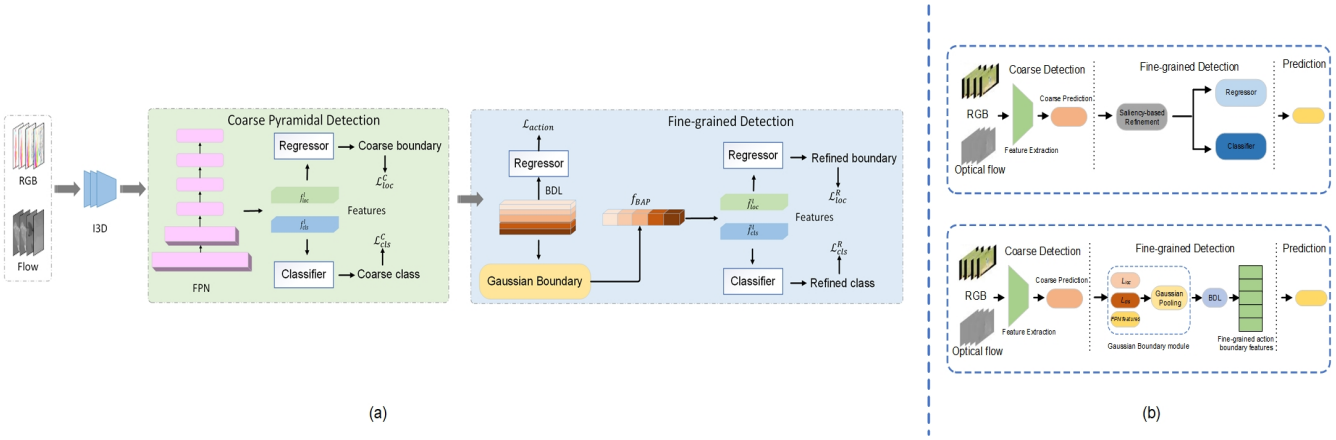


Figure 1. Architecture of the boundary awareness network (BAN) and its comparison with the AFSD. Panel (**a**) outlines the BAN structure, illustrating key components such as the Gaussian boundary module for precise boundary detection. Panel (**b**) presents a side-by-side structural comparison with the AFSD, highlighting the distinct features of the BAN, including Boundary Differentiated Learning (BDL) which is activated during training. This comparison delineates the specific architectural differences, demonstrating the BAN's approach to achieving more accurate action delineation.

3.1. Problem Definition

In the context of an unedited video V, comprising T frames, we segment it into n sections, each containing a fixed frame count. These segments are represented as

$$X = \{x_n\}_{n=1}^{l_v} \tag{1}$$

wherein x_n symbolizes the n-th frame within X. The temporal labeling set of X is constituted by a collection of temporal action instances, defined as

$$\Psi_g = \{\varphi_n = (t_{s,n}, t_{e,n})\}_{n=1}^{N_g} \tag{2}$$

with N_g denoting the total number of actual action instances, $t_{s,n}$ being the start time of the action instance φ_n, and $t_{e,n}$ representing its end time. Given these annotations for video V, our objective with the model is to accurately predict proposals, complete with class scores, ensuring both high recall and precision.

3.2. Boundary Awareness Network

The architecture of the proposed boundary awareness network is illustrated in Figure 1. The detection process of the BAN consists of three steps, including feature encoding, coarse pyramidal detection, and fine-grained detection.

Feature Encoding. In our approach, we have opted for the I3D [17] architecture as backbone, leveraging both RGB and optical flow frames as inputs. This choice is based on I3D's proven efficacy in both action recognition and temporal action localization tasks. Specifically, for a given input video V, we initially derive clip-level features $F = \{f_i\}_{i=0}^{T-1} \in \mathbb{R}^{T\times C\times H\times W}$ from consecutive clips. Here, T, C, H, and W represent the time step, channel, height, and width, respectively. We then compress the features along the last three dimensions into a one-dimensional feature sequence $F \in \mathbb{R}^{T\times C}$, encapsulating both the appearance and motion aspects of the video. In our study, the outputs from the final two layers are utilized as the predictive features.

Coarse Pyramidal Detection. To integrate the spatial dimension and aggregate the temporal dimension in multiple layers, we use a feature pyramid network with several temporal convolutions. Following recent methods [12], after feature encoding, we obtain a feature map $F \in \mathbb{R}^{T\times C}$ from raw video clips V, and then use $F \in \mathbb{R}^{T\times C}$ to generate FPN-level feature $F_{fpn} = f^l \in \mathbb{R}^{T^l\times c}$ with a series of convolutional layers, where l denotes a layer of the FPN level. The objective of the feature pyramid network (FPN) layer is to generate a representation of features across multiple scales, ensuring that each level of the feature maps possesses robust semantic details pertinent to temporal action localization. Then, we generate coarse prediction features by projecting these features into f_{cls} and f_{loc}, which contain action classification information and action position information, respectively. We can get coarse start and end boundary relative distances according to following formula:

$$d_i^{s,e} = F_t(f_{cls,loc}) \tag{3}$$

where F_t is one layer of temporal convolution shared among all FPN layers. Finally, we can calculate the absolute boundary position as follows:

$$\varphi_i = i * 2^l - d_i^s \tag{4}$$

$$\phi_i = i * 2^l - d_i^e \tag{5}$$

Subsequently, at each specific location, the predicted temporal regions are used to generate fine-grained features with Gaussian boundary pooling.

Fine-grained Detection. To amend the drawbacks of coarse pyramidal detection, we constructed a fine-grained feature via the Gaussian boundary module, as shown in Figure 2. Concatenating these features, we can get a boundary-awareness feature as follows:

$$f_{BAP} = \left(f, f_s^{'}, f_e^{'}, \hat{f}_s, \hat{f}_e\right) \tag{6}$$

where $f, f_s, f_e, \hat{f}_s, \hat{f}_e$ denote a frame-level feature, the start and end frame-level features, and start and end features, as detailed in Section 3.3. Additionally, we implemented Boundary Differentiated Learning (BDL) as a strategy to differentiate between action and non-action segments, effectively increasing the distinction between the characteristic features of action sequences and the background, as detailed in Section 3.4.

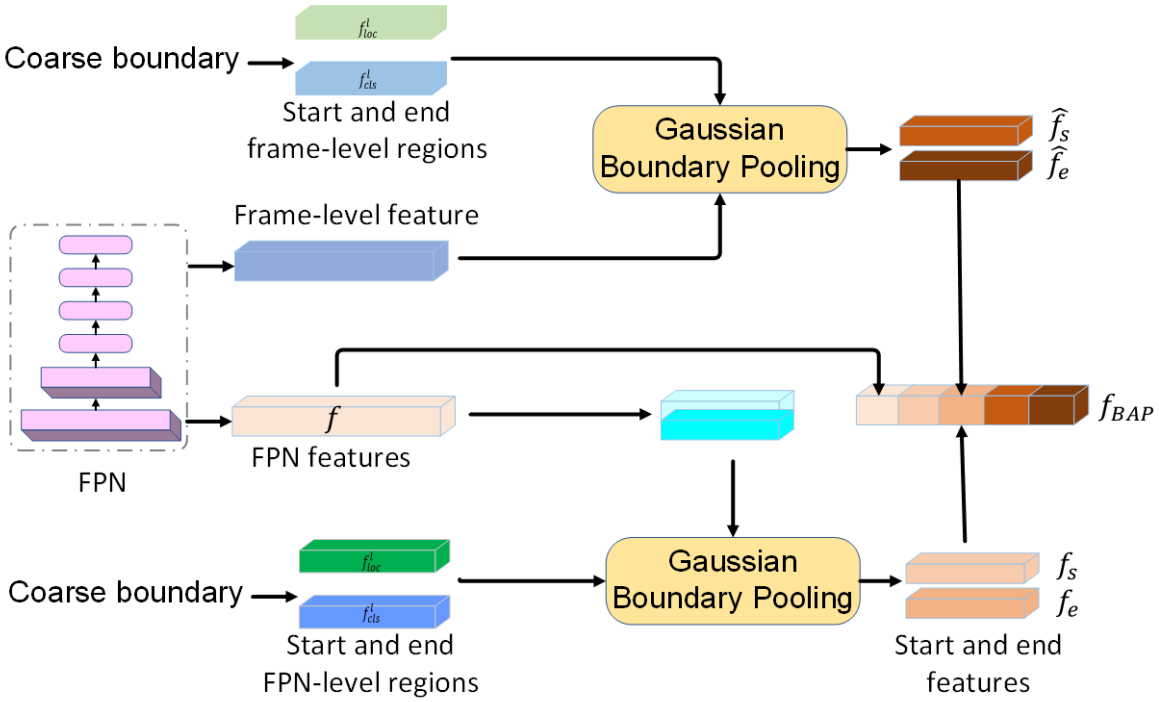


Figure 2. An illustration of the Gaussian boundary module.

3.3. Gaussian Boundary Module

The accurate localization of action boundaries within videos is a paramount challenge in temporal action localization, primarily due to the inherent ambiguity of action boundaries. The Gaussian boundary module (GBM) is designed to tackle this challenge by aggregating a comprehensive range of features to extract highly discriminative boundary features for accurate action prediction. These features encompass frame features along with start and end features, which are crucial for delineating the precise extents of actions.

The cornerstone of the GBM is the innovative use of Gaussian boundary pooling, a technique that significantly differs from traditional convolutional methods. This choice is motivated by our dual detection network structure, which incorporates both coarse-grained and fine-grained detection mechanisms. In this context, the coarse-grained detection outcomes serve not only as preliminary action proposals but also as critical guides for the subsequent fine-grained detection phase. This synergy is essential for refining the detection process and achieving high precision in boundary localization.

Gaussian boundary pooling operates by applying a Gaussian kernel centered on the predicted action boundaries from the coarse detection phase. The kernel's function is to weight the features according to their spatial–temporal proximity to the center, with closer features receiving higher weights. This method ensures that the boundary features are not just moment-level snapshots, but rather encompass a segment range, capturing the

nuanced transitions that characterize the start and end of actions. The Gaussian kernel's unique property of emphasizing features near the predicted boundary while smoothly decreasing the influence of distant features allows the GBM to extract boundary features that are both highly localized and contextually enriched.

This approach is particularly effective in capturing discriminative features that are vital for distinguishing between closely spaced actions and for accurately identifying action boundaries amidst complex background activities. To address this issue, we propose a novel differentiable pooling method called Gaussian boundary pooling, which is defined as follows:

$$\hat{f}^s(k) = f_{frame} \cdot g(x,t) \tag{7}$$

where $\cdot$ is matrix multiplication. The function $g(x,t)$ is specified as follows:

$$g(x,t) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{\left\|P - S^T\right\|^2}{2\sigma^2}\right) \tag{8}$$

where $P \in \mathbb{N}^{T\times T}$ is an automatically generated continuous sequence, and $S \in \mathbb{R}^{C\times T}$ consist of the k-th start and end time stamp, which is the result calculated by the coarse pyramidal detection stage . We empirically set the standard deviation σ to 0.5.

It is important to note that, although GTAN [5] has a similar structure, they are fundamentally different. In terms of design, the Gaussian kernel module employed in our work is derived from course results and serves the function of extracting the moment-range boundary feature. Furthermore, GTAN [5] employs Gaussian kernels to characterize action proposals with varying durations.

3.4. Boundary Differentiated Learning

In the realm of action detection within continuous video streams, the precision in delineating action boundaries is a nuanced challenge due to the inherently fuzzy nature of these boundaries. Traditional approaches often rely on discrete feature representation for action-versus-background discrimination, inadvertently leading the model to emphasize features that might not be critical for distinguishing subtle variations in action boundaries. This issue stems from the binary classification framework, which inherently lacks the granularity to accommodate the continuous variation observed in the temporal dynamics of actions.

To address this challenge and enhance the model's sensitivity to the continuum of action boundary features, we introduced Boundary Differentiated Learning (BDL). At its core, BDL re-envisions the task of action boundary detection by transitioning from a binary classification paradigm to a regression framework. This shift allows for a more nuanced representation of boundary features, reflecting the continuous nature of actions and their transitions.

The BDL strategy employs a novel approach to label assignment, where action boundaries are treated as a spectrum rather than discrete points. Specifically, we reformulated the label assignment process by introducing a continuous label space for the boundary regions. This is achieved through a re-scaling function applied to the ground truth labels, effectively transforming the binary classification task into a regression problem. The re-scaling function is designed to gradually adjust the label values based on their proximity to the action boundaries, thereby creating a gradient of importance that mirrors the fuzzy nature of these boundaries in real-world scenarios. The formula for re-measured boundary labels is as follows:

$$I_A(x) = \begin{cases} 1 & if \quad x \in Action \\ f(x) & if \quad x \in Boundary \\ 0 & if \quad x \in Background \end{cases} \tag{9}$$

where set $Boundary \in \mathbb{R}^l$ contains the predefined boundary area, and $f(x)$ is a re-measure function:

$$f(x) = 1 - \left(\frac{1}{d_{action} + x}\right) \tag{10}$$

where d_{action} is the length of the action, and x represents the position of the label. After such a re-measurement process, the boundary between action and background is no longer clear-cut but slowly changing. During model training, such a regression task enables the model to extract more robust boundary features. Following this, we proceed to compute the cross entropy:

$$L_{action} = CE_{loss}(\hat{j}, f_{model}) \tag{11}$$

where CE denotes the cross entropy loss function. In addition, to further improve the model accuracy and to facilitate the comparison with the baseline, we also used the Boundary Consistency Learning [12] strategy.

The incorporation of BDL into our model framework significantly bolsters its capability to discern and precisely localize action boundaries, even in challenging scenarios characterized by ambiguous or subtle transitions. By acknowledging and effectively leveraging the continuous nature of action boundaries, BDL adds substantial value to our model, enhancing its overall performance in temporal action localization tasks. This refined learning strategy, coupled with our model's architecture, marks a significant advancement in the field, pushing the boundaries of what is achievable in action detection within continuous video streams.

3.5. Training and Inference

Label Assignment. Following the AFSD [12], we evaluate the temporal Intersection over Union (tIoU) between each initial boundary prediction and its respective ground truth. A specific location i is considered positive if its tIoU score exceeds 0.5. Let N_C and N_R represent the count of positive samples for the initial and refined predictions, respectively. For the initial prediction, we label each location i as a positive sample corresponding to ground truth j if the condition $\varphi_i \leq i \leq \phi_j$ is met. Moreover, within the bounding region, we allocate position i to the recalibrated ground truth $\hat{j}$.

Training Objective. The BAN was trained using a composite multitask loss function, which encompasses the losses for coarse detection, fine-grained detection, action classification, and an L2 regularization term:

$$L_{total} = L_{coarse} + L_{fine-grained} + \gamma L(\Theta) \tag{12}$$

where γ is the hyperparameters. γ is set to 0.001.

Loss of coarse detection. In coarse detection, we directly process the FPN-level features with convolution to obtain the class of the action and the boundary time stamp. The optimization of the coarse detection can be achieved using the subsequent objective function:

$$L_{coarse} = \lambda_1 L_{cls} + \lambda_2 L_{loc} \tag{13}$$

where $lambda_1$ is set to 1, and $lambda_2$ is set to 10 and 1 on THUMOS14 and ActivityNet-1.3. L_{cls} is a softmax focal loss. We represent tIoU loss as L_{loc}.

Loss of fine-grained detection. In fine-grained detection, the loss function can be optimized through the application of the following specific function:

$$L_{fine-grained} = \lambda_1 L^f_{cls} + \lambda_2 L^f_{loc} + L_{BCL} + \gamma_1 L_{action} \tag{14}$$

where $lambda_1$ and $lambda_2$ settings are the same as in coarse detection. γ is set to 10 and 1 on THUMOS14 and ActivityNet-1.3, separately. The composition of L_{BCL} is consistent with that described in the AFSD [12]. We represent L1 loss as L^f_{loc}.

Inference. In the prediction phase, the BAN is utilized as outlined in the subsequent methodology:

$$\varphi_{start}^{coarse} = \varphi_{start}^{coarse} + \frac{1}{2} d_{coarse} \cdot \varphi_{start}^{fine-grained} \tag{15}$$

where φ_{start}^{coarse} denotes the action start time, which comes from the result of coarse-grained prediction. $\varphi_{start}^{fine-grained}$ denotes the fine-grained action start time. The end time is expressed in the same way as the start time. Subsequently, we compile all the predictions and apply Soft-NMS [20] for the elimination of overlapping proposals.

4. Experiments

4.1. Datasets

THUMOS14 [21] and ActivityNet1.3 [22] were adopted to demonstrate the performance of our BAN against numerous SOTA methods.

THUMOS14 [21]. The dataset includes 1010 videos in the validation set and 1574 in the testing set, spanning across 20 distinct categories. For temporal action localization specifically, it features 200 untrimmed videos in the validation subset and 212 in the testing subset, each marked with temporal annotations. Our model was trained using the validation subset and assessed on the test subset.

ActivityNet-1.3 [22]. For our purposes, we utilized version 1.3 of the dataset, which comprises 10,024 training videos, 4926 validation videos, and 5044 testing videos, each encompassing 200 distinct activity types. The majority of these videos feature instances of single-class activities occupying a substantial portion of their duration. In comparison to THUMOS14, it stands out as a more extensive dataset, both in the diversity of activities and the total volume of video content. The distribution ratio of the training, validation, and testing sets adheres to a 2:1:1 proportion.

4.2. Implementation Details

In our work with the THUMOS14 dataset [21], video encoding was performed at a rate of 10 frames per second (fps) and a spatial resolution of 96 $\times$ 96. In line with Lin's 2021 methodology, consecutive video clips were produced using sliding windows. Each clip was set to a temporal length of 256 frames, with an overlapping of 30 frames in the training phase and 128 frames during testing. For the ActivityNet1.3 dataset [22], we adjusted the frame sampling rates to guarantee that each video contained a total of 768 frames. As a result, this led to each video being represented by a single clip, consisting of 768 frames. The frame resolution was 96 $\times$ 96, which is the same as the size of THUMOS14. In training, data augmentation strategies such as random cropping and horizontal flipping were utilized. For feature extraction, we used the I3D [17] as a backbone to implement our framework, which was initialized with the parameters pre-trained on the Kinetics [23] dataset. Our experiment utilized a modified version of the ASFD model's code [12], specifically tailored to support our advanced boundary detection and learning features, while maintaining a comparative baseline with the original model.

Our approach was implemented in Python-3.8 using Pytorch on a GPU server with NVIDIA GeForce 1080Ti GPU(NVIDIA Corporation, Santa Clara, CA, USA). The total training epoch was set as 16 for both THUMOS14 and ActivityNet1.3. We utilized the Adam [24] optimizer with a learning rate set at 10^{-5} and a weight decay of 10^{-3}. Across various experiments, we consistently used a batch size of 1. During the training phase, the loss weight λ was adjusted to 10 for THUMOS14 and 1 for ActivityNet1.3, while γ was fixed at 10. In the testing phase, we averaged the results from both RGB and optical flow frames to determine the final locations and class scores. The temporal Intersection over Union (tIoU) threshold for Soft-NMS was set at 0.5 for THUMOS14 and 0.85 for ActivityNet1.3.

4.3. Metrics

For the task of temporal action detection, the mean Average Precision (mAP) served as the primary evaluation metric. In the case of ActivityNet-1.3, we utilized mAP at IoU

thresholds of 0.5, 0.75, and 0.95, along with an average mAP calculated over a range of IoU thresholds from 0.5 to 0.95 in increments of 0.05. For THUMOS14, the mAP was assessed at IoU thresholds of 0.3, 0.4, 0.5, 0.6, and 0.7.

4.4. Main Results

Our model was benchmarked against the leading methods in the field, as shown in Tables 1 and 2, across two demanding datasets, with a detailed report on the specifics employed by each approach, including BSN [8], GTAN [5], BMN [9], and the AFSD [12]. In the evaluation of our boundary awareness network (BAN) for temporal action localization, we conducted an extensive comparison with existing state-of-the-art methods, with a particular focus on the AFSD method, recognized for its leading performance in the domain.

Table 1. Comparison with state-of-the-art methods on THUMOS14 testing set.

Type	Model	Backbone	0.3	0.4	0.5	0.6	0.7	Avg.
Anchor-based	SSAD [1]	TS	43.0	35.0	24.6	-	-	-
	TURN [2]	C3D	44.1	34.9	25.6	-	-	-
	RC3D [3]	C3D	44.8	35.6	28.9	-	-	-
	CBR [25]	TS	50.1	41.3	31.0	19.1	9.9	30.3
	TAL [4]	I3D	53.2	48.5	42.8	33.8	20.8	39.8
	GTAN [5]	P3D	57.8	47.2	38.8	-	-	-
	PBRNet [6]	I3D	58.5	54.6	51.3	41.8	29.5	47.1
Actionness	CDC [11]	-	40.1	29.4	23.3	13.1	7.9	22.8
	SSN [7]	TS	51.0	41.0	29.8	-	-	-
	BSN [8]	TS	53.5	45.0	36.9	28.4	20.0	36.8
	BMN [9]	TS	56.0	47.4	28.8	29.7	20.5	38.5
	DBG [10]	TS	57.8	49.4	42.8	33.8	21.7	41.1
	GTAD [26]	TS	54.5	47.6	40.2	30.8	23.4	39.3
	BUTAL [27]	I3D	53.9	50.7	45.4	38.0	28.5	43.3
	BCGNN [28]	TS	57.1	49.1	40.4	31.2	23.1	40.2
Anchor-free	TadTR [29]	I3D	62.4	57.4	49.2	37.8	26.3	46.6
	AFSD [12]	I3D	67.3	62.4	55.5	43.7	31.1	52.0
	BAN (ours)	I3D	**68.8**	**63.9**	**56.4**	**46.1**	**32.7**	**53.6**

Table 2. Comparison with state-of-the-art methods on ActivityNet-1.3 testing set.

Type	Model	Backbone	0.5	0.75	Avg.
Anchor-based	SSAD [1]	TS	44.39	18.30	29.17
	TAL [4]	I3D	38.2	18.3	20.2
	GTAN [5]	P3D	52.6	34.1	34.3
Anchor–Actionness	CDC [11]	-	45.3	26.0	23.8
	SSN [7]	TS	43.2	28.7	28.3
	BSN [8]	TS	46.5	30.0	30.0
	BMN [9]	TS	50.1	34.8	33.9
	GTAD [27]	TS	50.4	34.6	34.1
	BCGNN [28]	TS	50.6	34.8	34.3
Anchor-free	TadTR [29]	I3D	49.1	32.6	32.3
	AFSD [12]	I3D	52.4	35.3	34.4
	BAN (ours)	I3D	52.5	35.5	34.6

Our BAN method introduced a novel Gaussian boundary module (GBM) and Boundary Differentiated Learning (BDL) strategy, specifically designed to enhance the detection of action boundaries in video sequences. This approach demonstrated superior performance on the challenging benchmark dataset THUMOS14 and showed promising results on ActivityNet-1.3, indicating its effectiveness across different contexts.

On the THUMOS14 dataset, the BAN achieved a mean Average Precision (mAP) of 56.4% at an IoU threshold of 0.5, outperforming the AFSD method, which recorded an mAP of 55.5%. This improvement is even more pronounced at higher IoU thresholds. For instance, at an IoU threshold of 0.7, the BAN achieved an mAP of 32.7%, which is a significant increase from the AFSD's 31.1%. Such results highlight the BAN's enhanced ability to accurately identify action boundaries, even in more stringent evaluation settings.

For ActivityNet v1.3, as detailed in Table 2, the BAN demonstrates a competitive performance with an mAP of 52.5% at an IoU threshold of 0.5 and shows significant improvement in more challenging conditions, with an mAP of 35.5% at an IoU threshold of 0.75 and an average mAP of 34.6%. These metrics not only affirm the model's robustness across varied datasets but also underscore its capacity to excel in identifying precise action boundaries amidst diverse and complex video contents.

The advantages of the BAN are attributed to its GBM, which effectively aggregates features around action boundaries, providing a richer and more discriminative representation of boundary features. Moreover, the BAN's Boundary Differentiated Learning (BDL) strategy facilitates the model in distinguishing between action and non-action segments more effectively, enhancing the overall precision of action localization.

In conclusion, the comparative analysis underscores the BAN's advancements over existing methods, including the AFSD. By focusing on the intricate detection of action boundaries and employing a refined learning strategy, the BAN sets a new benchmark for temporal action localization, promising significant improvements for real-world applications.

4.5. Ablation Study

We conducted multiple ablation tests on THUMOS14 to demonstrate the impact of key components of the BAN. In this section, we present experiments with the goal of identifying the main elements that contribute to high-quality temporal action detection.

The Effect of the Gaussian Boundary Module. To ascertain the specific contribution of the Gaussian boundary module (GBM) to temporal action localization (TAL), we devised a comprehensive series of ablation experiments. These experiments were designed to evaluate two primary aspects: (1) the impact of varying the range of boundary features on detection accuracy, and (2) the efficacy of different Gaussian boundary pooling methods within the GBM.

The results, detailed in Table 3, illustrate a clear trend: as the boundary feature range narrows (indicated by a decrease in σ), the localization accuracy significantly improves, peaking at $\sigma = 0.5$. This parameter, σ, represents the variance within the Gaussian function, with lower values indicating a more focused aggregation of features around the action boundaries. The optimal performance at $\sigma = 0.5$ underscores the necessity for boundary features to be scope-specific, enabling the precise delineation of action start and end points, as opposed to being broadly defined or moment-specific.

Table 3. The effect of the range of the boundary on the detection. The score '0.5' used in our model gets the highest performance.

σ	0.3	0.4	0.5	0.6	0.7	Avg.
5	58.8	53.1	49.2	39.8	24.6	45.1
1	65.1	59.9	55.2	45.4	30.9	51.3
0.5	68.8	63.9	56.4	46.1	32.7	53.6
0.2	67.2	62.4	55.1	45.9	32.3	52.0

Impact of Boundary Differentiated Learning (BDL). Further, to quantify the influence of Boundary Differentiated Learning (BDL) on the model's performance, we compared the BAN's effectiveness with and without the implementation of BDL. This comparison revealed that BDL significantly enhances the model's ability to distinguish between action and non-action segments, thereby improving the overall accuracy of action localization. The integration of BDL into the BAN led to a marked increase in mean Average Precision

(mAP) across various IoU thresholds, particularly highlighting its role in refining the model's predictive capabilities in complex video sequences where action boundaries may be less distinct.

Combined Influence of the GBM and BDL. In addition to evaluating the individual contributions of the GBM and BDL, our study also examined their combined effect on the model's performance. The full incorporation of both the GBM and BDL into the BAN resulted in the highest observed improvements in localization accuracy, demonstrating the synergistic impact of these components. This combined configuration significantly outperformed other variations, including models with either component omitted, thereby validating the integral role of both the GBM and BDL in enhancing the efficacy of the BAN for TAL.

The findings from these ablation studies are encapsulated in Table 4, which presents a detailed breakdown of performance metrics under various configurations of the BAN model. Through this rigorous analysis, we have established the critical contributions of both the GBM and BDL to the overall performance of the BAN, affirming their value in our proposed framework for temporal action localization.

Table 4. The effect of the range of the Gaussian boundary module. The second and third columns represent different levels of features for Gaussian boundary pooling, respectively.

	Frame-Level	FPN-Level	0.3	0.4	0.5	0.6	0.7	Avg.
BAN without GBM			61.2	58.4	49.8	38.7	25.5	46.7
BAN with GBM			66.6	61.0	53.8	44.5	30.9	49.1
	√		68.2	62.8	55.8	46.0	32.5	53.1
		√	68.1	63.4	55.6	45.8	32.3	53.0
	√	√	68.8	63.9	56.4	46.1	32.7	53.6

4.6. Comparison of Inference Speed

In this section, we compare the detection speeds of our model with two other leading methods. The inference speed on the THUMOS14 dataset was assessed in terms of frames per second (fps) across various models, as outlined in Table 5. RGB frames were employed as inputs for all networks to ensure a balanced comparison. Our findings indicate that our model operates significantly faster than most existing methods, with only a marginal difference compared to the benchmark model AFSD. Specifically, the execution speed of our model was recorded at 3197 fps on a single RTX-1080ti (NVIDIA Corporation, Santa Clara, CA, USA) GPU for temporal action localization tasks, which escalated to an impressive 4291 fps on the more advanced RTX-2080ti (NVIDIA Corporation, Santa Clara, CA, USA) GPU.

Table 5. Activity detection speed during inference.

Method	SS-TAD	R-C3D	AFSD	PBR	BAN (Ours)	
GPU	TITAN Xm	TITAN Xp	1080Ti	1080Ti	1080Ti	2080Ti
FPS	701	1030	3259	1488	3197	4291

While our model demonstrated a high inference speed, making it suitable for near-real-time applications, we acknowledge the ongoing need to enhance its efficiency for real-time processing in live video streams. Future optimizations may include algorithmic improvements, model simplification, and leveraging edge computing technologies to further reduce latency and computational requirements. These efforts aim to expand the applicability of our model to real-time scenarios, ensuring it can deliver a high performance in a wider range of practical applications.

4.7. Visualization Results

In this section, we compare the visualization results of our method and the state-of-the-art method AFSD [12] on the THUMOS14 test set and the results are shown in Figure 3. Our proposed method has a greater ability to find action boundaries, according to the results.

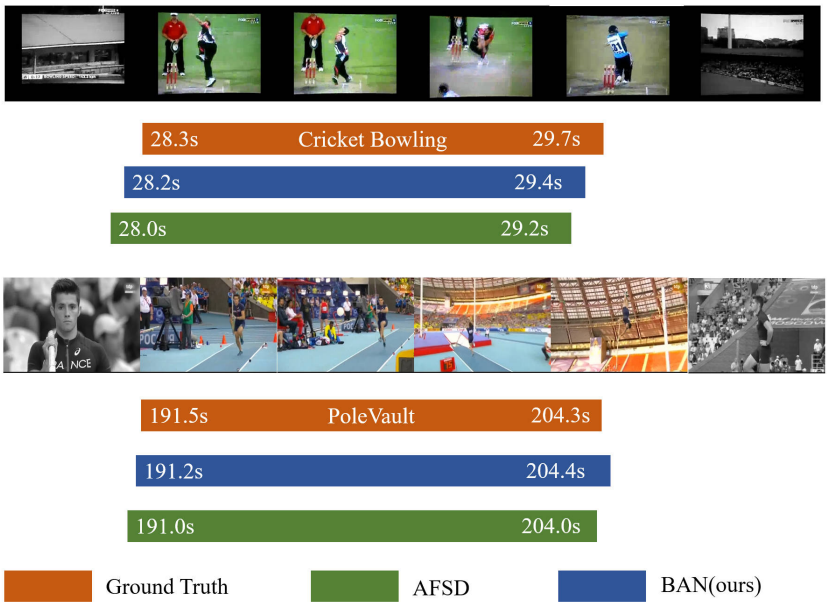


Figure 3. Visualization results of our method and AFSD on THUMOS14.

5. Conclusions

In this study, we introduced a novel action localization framework named BAN, which achieves state-of-the-art performance on standard benchmarks. Our method generates precise boundary proposals by aggregating features around the boundary and refines these predictions to make more accurate boundary determinations. We proposed the Gaussian boundary module (GBM) to obtain accurate boundary features, coupled with a novel Boundary Differentiated Learning (BDL) strategy to discern discriminative action information effectively. Extensive experiments conducted on well-known benchmarks, including ActivityNet-1.3 and THUMOS14, underscore the BAN's advanced capabilities in comparison with existing state-of-the-art methods, including the AFSD. This contribution significantly enriches the community's toolbox for video understanding, offering a powerful means to tackle the nuanced challenges of action localization with unprecedented precision. Moreover, by setting new benchmarks, the BAN paves the way for future research to build upon our findings, fostering innovation and further exploration in the realm of video analysis.

Acknowledging the limitations of our approach is crucial for setting realistic expectations and guiding future work; while the BAN represents a significant advancement in temporal action localization, its performance is contingent upon the quality of feature representations and may encounter challenges in cluttered scenes where actions are less distinct. These limitations highlight the importance of ongoing research efforts to enhance the robustness of the BAN, particularly in complex scenarios. Future directions include exploring advanced techniques for feature extraction and representation, aiming to improve the model's adaptability and effectiveness across a broader range of video content. Moreover, addressing these challenges will involve developing methods that can more effectively navigate the complexities of video scenes, thereby ensuring the BAN's relevance and applicability as video analysis challenges evolve. To further improve performance in complex scenes, we plan to explore the potential of integrating context-aware models, such as scene segmentation or object recognition. These models could provide richer contextual cues for action boundary detection, thereby improving action localization accuracy in

visually complex scenes. By introducing this level of context understanding, we anticipate future iterations of our framework to better handle the subtle distinctions between actions and backgrounds, offering deeper insights for video analysis. Additionally, we aim to incorporate more advanced deep learning techniques, like attention mechanisms or transformers, in future iterations to more accurately capture the temporal dynamics of videos, further enhancing our model's ability to discern and localize actions within complex video sequences. To address the dependence on quality feature representations, we also intend to explore unsupervised or semi-supervised learning methods in future work. These approaches could enhance the model's ability to learn from unlabeled data, potentially reducing the reliance on extensively annotated datasets and improving the model's generalization capabilities.

In light of the advancements in temporal action localization (TAL) demonstrated by our model, it is crucial to acknowledge the ethical implications associated with its application. The potential for enhancing understanding and interaction with video content comes with a responsibility to consider privacy concerns and guard against misuse. As TAL technologies become increasingly capable, their application in surveillance, media analysis, and beyond must be approached with a commitment to ethical standards and privacy protection. We advocate for ongoing dialogue and the establishment of robust guidelines to ensure that the benefits of TAL are realized in a manner that respects individual rights and societal values.

Author Contributions: J.G.: As the first author, J.G. was instrumental in the conceptualization and design of the study. He was primarily responsible for the development and execution of the coding and experimental work. Additionally, J.G. contributed to part of the manuscript writing, ensuring that the experimental results and methodologies were accurately and clearly presented. Y.Y.: As the second author, Y.Y. provided comprehensive guidance throughout the project. Her role was pivotal in overseeing the research direction and methodology. Y.Y.'s extensive experience and knowledge were crucial in formulating the research questions and in interpreting the results. She also provided significant input during the manuscript revision stages, ensuring the overall quality and integrity of the research. M.W.: The third author, M.W., played a significant role in the manuscript writing process. She actively participated in drafting the manuscript and revising it critically for important intellectual content. Her contributions also included engaging in discussions and providing insights that significantly shaped the research and analysis. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partly supported by National Natural Science Foundation of China (No. 61672546), and Guangzhou Science and Technology Project (No. 201707010127).

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Acknowledgments: The authors would like to thank the editor and anonymous reviewers for their contributions towards improving the quality of this paper.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Lin, T.; Zhao, X.; Shou, Z. Single shot temporal action detection. In Proceedings of the 25th ACM International Conference on Multimedia, Mountain View, CA, USA, 23–27 October 2017; pp. 988–996.
2. Gao, J.; Yang, Z.; Chen, K.; Sun, C.; Nevatia, R. Turn tap: Temporal unit regression network for temporal action proposals. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3628–3636.
3. Xu, H.; Das, A.; Saenko, K. R-C3D: Region convolutional 3D network for temporal activity detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5783–5792.
4. Chao, Y.-W.; Vijayanarasimhan, S.; Seybold, B.; Ross, D.A.; Deng, J.; Sukthankar, R. Rethinking the Faster R-CNN architecture for temporal action localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1130–1139.
5. Long, F.; Yao, T.; Qiu, Z.; Tian, X.; Luo, J.; Mei, T. Gaussian temporal awareness networks for action localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 344–353.

6. Liu, Q.; Wang, Z. Progressive boundary refinement network for temporal action detection. *Proc. Aaai Conf. Artif. Intell.* **2020**, *34*, 11612–11619. [CrossRef]
7. Zhao, Y.; Xiong, Y.; Wang, L.; Wu, Z.; Tang, X.; Lin, D. Temporal action detection with structured segment networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2914–2923.
8. Lin, T.; Zhao, X.; Su, H.; Wang, C.; Yang, M. BSN: Boundary sensitive network for temporal action proposal generation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
9. Lin, T.; Liu, X.; Li, X.; Ding, E.; Wen, S. BMN: Boundary-matching network for temporal action proposal generation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3889–3898.
10. Lin, C.; Li, J.; Wang, Y.; Tai, Y.; Luo, D.; Cui, Z.; Wang, C.; Li, J.; Huang, F.; Ji, R. Fast learning of temporal action proposal via dense boundary generator. *Proc. Aaai Conf. Artif. Intell.* **2020**, *34*, 11499–11506. [CrossRef]
11. Shou, Z.; Chan, J.; Zareian, A.; Miyazawa, K.; Chang, S.-F. CDC: Convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5734–5743.
12. Lin, C.; Xu, C.; Luo, D.; Wang, Y.; Tai, Y.; Wang, C.; Li, J.; Huang, F.; Fu, Y. Learning salient boundary feature for anchor-free temporal action localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 3320–3329.
13. Vaudaux-Ruth, G.; Chan-Hon-Tong, A.; Achard, C. SALAD: Self-Assessment Learning for Action Detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Virtual, 5–9 January 2021; pp. 1269–1278.
14. Law, H.; Deng, J. CornerNet: Detecting Objects as Paired Keypoints. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 734–750.
15. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: Fully Convolutional One-Stage Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9627–9636.
16. Qiu, H.; Ma, Y.; Li, Z.; Liu, S.; Sun, J. BorderDet: Border Feature for Dense Object Detection. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 549–564.
17. Carreira, J.; Zisserman, A. Quo vadis, action recognition? A new model and the kinetics dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6299–6308.
18. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*. [CrossRef] [PubMed]
19. Yang, L.; Peng, H.; Zhang, D.; Fu, J.; Han, J. Revisiting Anchor Mechanisms for Temporal Action Localization. *IEEE Trans. Image Process.* **2020**, *29*, 8535–8548. [CrossRef]
20. Bodla, N.; Singh, B.; Chellappa, R.; Davis, L.S. Soft-NMS–Improving Object Detection with One Line of Code. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5561–5569.
21. Idrees, H.; Zamir, A.R.; Jiang, Y.-G.; Gorban, A.; Laptev, I.; Sukthankar, R.; Shah, M. The THUMOS challenge on action recognition for videos "in the wild". *Comput. Vis. Image Underst.* **2017**, *155*, 1–23.
22. Caba Heilbron, F.; Escorcia, V.; Ghanem, B.; Carlos Niebles, J. ActivityNet: A large-scale video benchmark for human activity understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 961–970.
23. Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; et al. The kinetics human action video dataset. *arXiv* **2017**, arXiv:1705.06950.
24. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
25. Gao, J.; Yang, Z.; Nevatia, R. Cascaded boundary regression for temporal action detection. *arXiv* **2017**, arXiv:1705.01180.
26. Xu, M.; Zhao, C.; Rojas, D.S.; Thabet, A.; Ghanem, B. G-TAD: Sub-graph localization for temporal action detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10156–10165.
27. Zhao, P.; Xie, L.; Ju, C.; Zhang, Y.; Wang, Y.; Tian, Q. Bottom-up temporal action localization with mutual regularization. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 539–555.
28. Bai, Y.; Wang, Y.; Tong, Y.; Yang, Y.; Liu, Q.; Liu, J. Boundary content graph neural network for temporal action proposal generation. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 121–137.
29. Liu, X.; Wang, Q.; Hu, Y.; Tang, X.; Bai, S.; Bai, X. End-to-end temporal action detection with transformer. *arXiv* **2021**, arXiv:2106.10271.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.