Progressive Attention on Multi-Level Dense Difference Maps for Generic Event Boundary Detection

Jiaqi Tang[1] Zhaoyang Liu[2] Chen Qian[2] Wayne Wu[2] Limin Wang[1] ✉
[1]State Key Laboratory for Novel Software Technology, Nanjing University, China
[2]SenseTime Research
jqtang@smail.nju.edu.cn, zyliumy@gmail.com, {qianchen, wuwenyan}@sensetime.com, lmwang@nju.edu.cn

Abstract

Generic event boundary detection (GEBD) is an important yet challenging task in video understanding, which aims at detecting the moments where humans naturally perceive event boundaries. The main challenge of this task is perceiving various temporal variations of diverse event boundaries. To this end, this paper presents an effective and end-to-end learnable framework (DDM-Net). To tackle the diversity and complicated semantics of event boundaries, we make three notable improvements. First, we construct a feature bank to store multi-level features of space and time, prepared for difference calculation at multiple scales. Second, to alleviate inadequate temporal modeling of previous methods, we present dense difference maps (DDM) to comprehensively characterize the motion pattern. Finally, we exploit progressive attention on multi-level DDM to jointly aggregate appearance and motion clues. As a result, DDM-Net respectively achieves a significant boost of 14% and 8% on Kinetics-GEBD and TAPOS benchmark, and outperforms the top-1 winner solution of LOVEU Challenge@CVPR 2021 without bells and whistles. The state-of-the-art result demonstrates the effectiveness of richer motion representation and more sophisticated aggregation, in handling the diversity of GEBD. The code is made available at https://github.com/MCG-NJU/DDM.

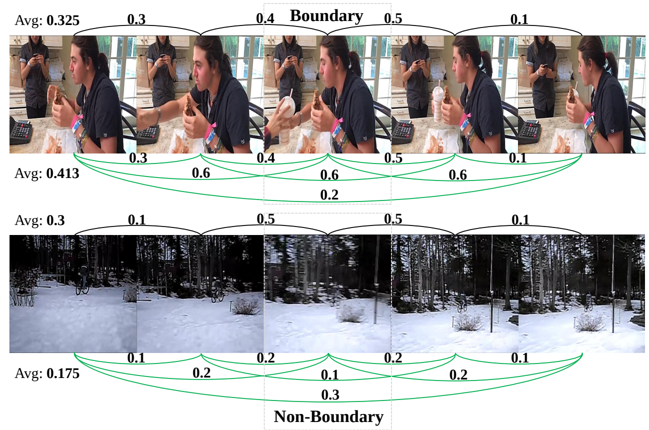

Figure 1. **Comparisons of sparse motion representation** (black lines, optical flow) **and dense motion representation** (green lines, some are omitted for clarity, dense feature differences). Numbers on lines indicate the magnitude of motion between two frames. Dense motion representation provides more holistic temporal cues to better distinguish boundaries and non-boundaries.

1. Introduction

With the explosive growth of online videos, video understanding has drawn tremendous attention from both academia and industry. Cognitive science [43] suggests that humans naturally divide a video into meaningful units by perceiving event boundaries. To this end, a task termed as **Generic Event Boundary Detection** [35] (GEBD) is recently proposed to localize the generic event boundaries in videos, which is expected to facilitate the development of video understanding.

Generic event boundaries in GEBD task are taxonomy-free and related to a broad range of temporal changes, including changes of action, subject and environment. The primary challenge in GEBD task is to model diverse patterns of generic event boundaries: *a) Spatial diversity* is dominantly characterized by the change of appearance, which normally comprises low-level changes (*e.g.*, change in color or brightness) and high-level changes (*e.g.*, the dominant subject appears or disappears). *b) Temporal diversity* is mainly relevant to actions, such as change of action (*e.g.*, walk to run) or change of object of interaction. Notably, different actions usually exhibit inconsistent speed and duration, which further increases the temporal diversities of event boundaries. As a result, the spatio-temporal diversities lead to overly complicated variations in videos, which impedes the accurate detection of event boundaries.

✉: Corresponding author.

Since GEBD task is highly correlated with changes in temporal dimension, motion information is the key to perceiving temporal variations and detecting event boundaries. Previous methods wildly use optical flow [24, 25, 45] as alternative motion representation to learn temporal clues in videos. However, they model the semantics in a single feature level and focus on local motion cues between two consecutive frames (Figure 1), which is insufficient to perceive diverse event boundaries. In addition, previous two-stream methods [36, 45] commonly resort to simple fusion schemes, short of interaction across appearance and motion modalities. Hence, they are less effective for learning complex semantics of diverse event boundaries.

To address the above issues, we present a method (DDM-Net) that progressively aggregates dense motion information along with appearance cues to perceive event boundaries, as illustrated in Figure 2. We make three notable improvements, including Multi-Level Feature Bank, Dense Difference Map and Progressive Attention. *First*, we build a *Multi-Level Feature Bank* where the features are collected in different spatial and temporal scales respectively, which empowers the subsequent modules to thoroughly perceive different levels of changes in videos.

Second, based on aforementioned feature bank, we propose a *Dense Difference Map* (DDM) to model rich temporal contexts. Technically, we calculate pairwise feature differences between every two frames in a clip of length T, and obtain a $T \times T$ dense difference map. The main advantage of DDM is to exploit the difference of each feature pair and provide holistic motion information. As shown in Figure 1, our proposed DDM is able to provide more holistic and salient temporal clues than optical flow, which is calculated between two consecutive frames. Furthermore, instead of directly being operated on raw frames, our DDM is built on the features collected from different layers of backbone network, and thus ought to be more robust to temporal noise (*e.g.*, camera blur in the second row of Figure 1).

Third, as event boundaries show their spatio-temporal diversities and complexities, we argue that simple linear fusion in two-stream methods is insufficient to aggregate the appearance and motion clues. We thus exploit *Progressive Attention* to mine important clues hidden in RGB features and DDM. In order to align the shape of DDM to RGB features, we design map-squeezed attention to squeeze DDM. Then, in intra-modal attention, key features of two modalities are respectively enhanced through two sets of learnable queries, prepared for cross-modal attention. Cross-modal attention is leveraged to perform feature interaction across modalities, enabling appearance and motion features to query and guide each other. As a result, DDM-Net can more effectively aggregate spatio-temporal clues and improve the discrimination of event boundaries.

Our DDM-Net exploits multi-level dense differences to perceive diverse temporal variations, and leverages progressive attention to effectively aggregate appearance and motion clues. To prove the effectiveness of DDM-Net, we perform extensive experiments on two datasets: Kinetics-GEBD [35] and TAPOS [34]. Evaluation results demonstrate that our DDM-Net outperforms the existing state-of-the-art methods by a large margin on all evaluation metrics. Particularly, DDM-Net obtains a superior 76.4% F1@0.05 on Kinetics-GEBD, with a significant boost of 14 percent. On TAPOS, we improve F1 score@0.05 from 52.2% to 60.4%. In addition, our DDM-Net is superior to winners of LOVEU Challenge@CVPR 2021 [35] on the testing set of Kinetics-GEBD, demonstrating the effectiveness of our method. In summary, our main contributions are as follows:

- We propose dense difference maps equipped with multi-level feature bank to leverage richer temporal clues for detection of diverse event boundaries.
- Instead of simple feature fusion methods, progressive attention is employed to aggregate appearance and motion clues from RGB features and DDM, enabling DDM-Net to generate more discriminative representations and learn more complicated semantics.
- Extensive experiments and studies demonstrate that our DDM-Net achieves the state-of-the-art performance on Kinetics-GEBD and TAPOS benchmark, under the setting of the same backbone.

2. Related Work

Temporal Detection Tasks in Video Understanding. Temporal action detection task aims to detect action instances in untrimmed videos, namely predict starting point, ending point and category of each action. One-stage and two-stage methods are two mainstream solutions. Different from direct one-stage methods [3, 23, 47], two-stage methods [24, 25, 27, 33, 39, 49] decompose the task into class-agnostic proposal generation and action classification. Temporal action parsing [34] is recently proposed, of which the target is dividing actions into segments of sub-actions. Video anomaly detection [12, 26, 31, 37] is aimed at recognizing frames where abnormal events happen, wildly applied in video surveillance. As for shot boundary detection [2, 13, 40], it is a classical task for significant shot change detection. Different from them, GEBD [35] is a generic detection task, where generic event boundaries include all of the above. To address the diversity and complicated semantics of generic event boundaries, our method improves the boundary discrimination via progressively attending to multi-level dense difference maps.

Motion Representation. Previous methods of current video understanding tasks (*e.g.*, action recognition, temporal action detection, *etc.*) wildly used optical flow [4, 24, 36,

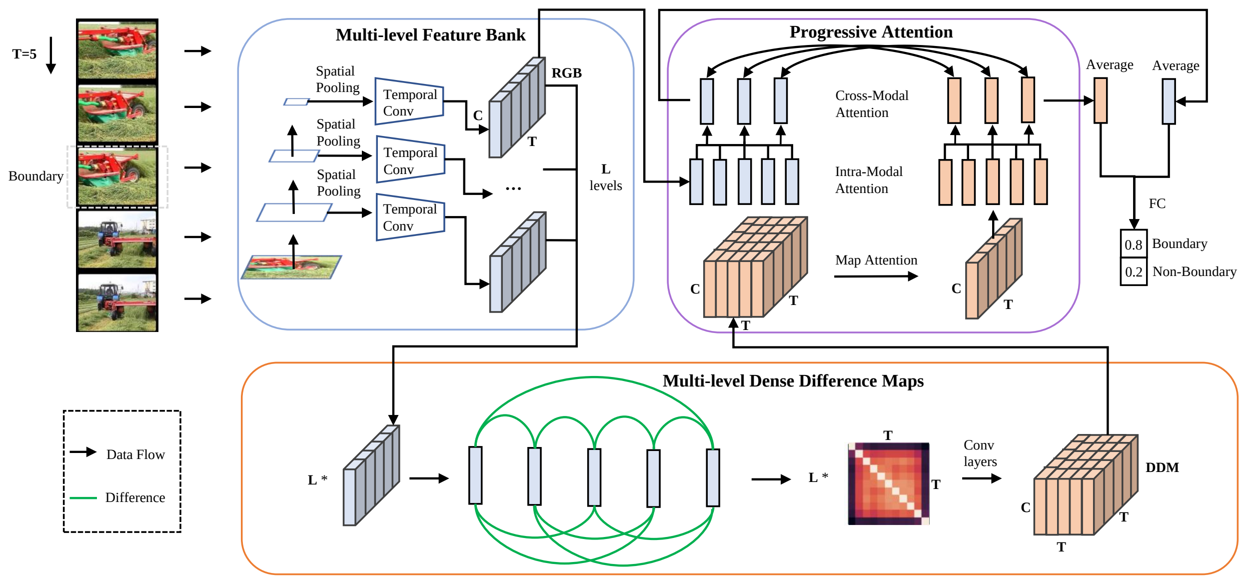

Figure 2. **Overview of DDM-Net.** Our DDM-Net streamlines the process of generic event boundary detection by viewing it as a binary classification problem of sliding video clips. Specifically, our method classifies the current frame with a clip centered on it and repeats the same process on other frames. The network is mainly composed of three stages: multi-level feature bank construction, dense difference map calculation, and progressive attention. DDM-Net exploits richer motion information and more sophisticated aggregation to achieve accurate detection for generic event boundaries. (L: number of levels of features, T: number of frames, C: number of channels.)

45], RGB differences [18, 32, 45, 51, 52] and feature differences [8, 17, 22, 28, 29, 44] as the motion representation to learn temporal clues in videos. However, they focus on local motion cues between two consecutive frames and do not explicitly employ multi-level features for complex semantics learning. Compared with single-level sparse motion representations, our proposed multi-level DDM is a dense motion representation built upon the multi-level feature bank, better to perceive diverse temporal variations in multiple levels.

Multi-Modal Feature Aggregation and Fusion. Multi-modal feature aggregation and fusion are wildly applied for holistic semantics learning in many tasks, such as image-text [1, 30, 46, 48], video-text [10, 14, 21] and audio-video tasks [11, 41]. In video understanding, previous two-stream methods [9, 36, 45] train two separate networks and resort to a simple fusion of two video modalities, namely appearance and motion, via linear fusion or feature concatenation. However, they are less effective due to a lack of interaction across two modalities, falling short of dependencies between appearance and motion features [6]. To this end, our method progressively attends to multi-level DDM, taking advantage of correlations between two modalities to enrich the semantic information and improve the discrimination.

3. Method

3.1. Overview

Generic Event Boundary Detection [35] (GEBD) aims to detect the taxonomy-free event boundaries, *e.g.*, change of action, change of subject, shot change, etc. As the temporal boundaries in video usually exhibit the dominant characteristics of ambiguity and diversity, it indeed is a challenging vision task remaining to be studied. To this end, we propose a novel spatio-temporal modeling scheme, which constructs and attends to multi-level dense difference maps to address aforementioned issues.

Given a video $V = \{I_t\}_{t=1}^{E}$, where I_t is the t-th frame and E is the number of frames in video, we sample a clip $U = \{I_{t-w}, ..., I_t, ..., I_{t+w}\}$ of T ($T = 2 \times w + 1$) frames from video V to infer whether I_t is a boundary frame. As illustrated in Figure 2, DDM-Net mainly refers to three parts: a multi-level spatio-temporal feature bank, multi-level dense difference maps and cross-modal aggregation between **RGB features** A and **Dense Difference Map** (DDM) M via progressive attention. *Firstly*, the sampled clip U is fed into a backbone network and a serial of temporal convolutions to yield *Multi-Level* spatio-temporal features $\mathbf{F} = \{f_{ij}\}_{i\in[1,m],j\in[1,n]}$, where i and j respectively denotes the spatial level (m levels in total) and the temporal level (n levels in total) of the feature. *Secondly*, a *Dense Difference Map* $M \in \mathbb{R}^{C\times T\times T}$ is constructed with $\mathbf{F}$ by measuring the discrepancy among frames, which is expected to provide more discriminative information to aid the model in perceiving temporal variations. *Thirdly*, in our *Progressive Attention* module, intra-modal attention module exploits a set of learnable queries to enhance key intra-modal representations, and co-attention transformers are leveraged to perform cross-modal attention. It is worth noting that, to align with RGB features A, DDM $M \in \mathbb{R}^{C\times T\times T}$ is firstly squeezed to a sequence $D \in \mathbb{R}^{C\times T}$ via map-squeezed attention. *Finally*, A and D are respectively fed into separate fully-connected (fc) layers. With a linear fusion after fc layers, the model outputs the final boundary probability of the center frame I_t.

In contrast to previous two-stream methods [9, 36, 45] that use optical flow as motion representation to learn temporal clues in videos, we meticulously construct dense difference maps, which enable the model to perceive generic event boundaries along with RGB features. Since DDM is calculated on-the-fly, our method is more efficient than previous two-stream methods that train two separate networks. In the following sections, we will introduce the technical details of each module.

3.2. Multi-Level Feature Bank

To model diverse motion patterns of generic event boundaries, we exploit a feature bank to store multi-level features of input video clips, based on which the dense difference map is calculated to yield rich temporal clues.

Temporal View of Multi-Level Feature Bank. Before building the feature bank, an issue that needs to be figured out is whether we take a clip or the whole video as inputs to detect event boundaries. As videos normally are composed of multiple non-overlapping and relatively independent snippets that belong to different events, we argue that whether the current frame is an event boundary is mostly related to its adjacent snippets. Snippets far away from the current frame contribute little to infer whether it is an event boundary. Therefore, we opt to build our model based on a clip around the current frame, instead of the whole video. Notably, experiments in Table 4b have also examined the rationality of our point. Specifically, along with the current frame, we sample w frames before and after the current frame, namely T ($T = 2 \times w + 1$) frames as an input clip. Then, the input clip is fed into backbone network to construct the multi-level feature bank.

Construction of Multi-Level Feature Bank. Since event boundaries in GEBD task are generic and taxonomy-free, patterns of different event boundaries vary considerably in space and time. From a perspective of space, appearance changes include low-level changes and high-level changes. Low-level changes mainly refer to change in environment (*i.e.*, change in color and brightness), while high-level changes are related to complex semantics (*e.g.*, the dominant subject appears or disappears). From a temporal perspective, the duration of action changes is usually inconsistent. For instance, 'a runner suddenly changes direction' can happen very fast, while 'an old man slowly stands up' usually takes several frames.

To detect event boundaries with diverse motion patterns, our method models temporal variations upon multi-level spatio-temporal features. Specifically, we perform average spatial pooling on m layers of ResNet features (*e.g.*, layer3 and layer4), and get m feature sequences of different semantic levels. It is notable that the feature sequence of high-level layer4 is also denoted as RGB features A, which are later fused with DDM features D, as shown in Figure 2. Then, for each feature sequence, we exploit temporal convolutions to get n feature sequences with different temporal receptive fields. Consequently, there are $m \times n = L$ levels of features in total, prepared for multi-level dense difference calculation. In Table 4c, we observe that both multi-level features from spatial and temporal domain provide crucial clues to detect diverse event boundaries.

3.3. Dense Difference Maps

Motion representation is crucial in GEBD task. As for boundaries like changes of action, there is little change in appearance (*e.g.*, a man waves gently towards the camera, or walk to run). To detect such boundaries, motion information plays a principal role in perceiving temporal variations. Previous methods commonly exploit sequential optical flow or RGB differences to approximate motion information. However, they can only reflect local motion cues between two consecutive frames and fail to take advantage of rich temporal contexts. Considering the variety of boundaries and complicated scenarios in GEBD task, it is insufficient to use local and sparse motion representation.

To alleviate inadequate temporal context modeling of sparse motion representation, we propose dense difference maps based on aforementioned multi-level feature bank. Given a feature sequence of T frames, we calculate the feature difference of each frame pair and construct a $T \times T$ map. Compared with the sparse motion sequence of length $T-1$, $T \times T$ pairs of feature differences provide denser temporal cues (Figure 1). Since DDM contains richer motion information, it characterizes the motion pattern around the current frame more holistically, enabling our method to better perceive temporal variations and distinguish boundaries and non-boundaries. Moreover, DDM is constructed with aforementioned multi-level feature bank, where features are collected from different layers of backbone network and consist of multi-level semantics. Hence, it is more robust to temporal noise than optical flow and RGB differences, which are directly calculated on raw frames.

In practice, Euclidean distance is employed across all the channels to measure the feature difference between two frames I_i and I_j,

$$FD(i,j) = \sqrt{\sum_{c=1}^{C}(A_i^c - A_j^c)^2}, \tag{1}$$

where A_i and A_j are appearance features of I_i and I_j, C is the total number of channels. Then, we exploit stacked convolution layers to transform difference matrices$\in \mathbb{R}^{L\times T\times T}$ into $M \in \mathbb{R}^{C\times T\times T}$. In Table 4d, DDM-Net also achieves close performance with other distance metrics(*e.g.*, Manhattan distance), which demonstrates the performance of our method is robust to the choice of difference operators.

3.4. Progressive Attention

Previous two-stream networks usually leverage simple aggregation and fusion manners, such as linear fusion or feature concatenation of temporal averaging results. However, they lack interaction between modalities and thus cannot take full advantage of our proposed DDM, which is proved in Table 4e. Hence, to better aggregate appearance and motion clues, we employ progressive attention on our proposed multi-level DDM, including map-squeezed attention, intra-modal attention and cross-modal attention.

Map-Squeezed Attention. To align $M \in \mathbb{R}^{C\times T\times T}$ with RGB features $A \in \mathbb{R}^{C\times T}$, we transform it into a feature sequence of length T via frame-wise map-squeezed attention. In DDM, feature sequence of the i-th row ($M_i \in \mathbb{R}^{C\times T}$) is the difference between the i-th frame I_i and other frames of the current clip. Hence, it is intuitive to aggregate elements of the M_i to get a clip-level motion measurement of the I_i. Due to the diversity of temporal dependencies, it is common that differences with several specific frames are more important than others. Therefore, we propose a frame-wise attention mechanism to squeeze M, calculating the weights of all elements in M_i based on feature A_i of I_i. Concretely, we exploit A_i to attend all elements of M_i and generate weights γ_i, adaptively aggregating all differences into a motion measurement D_i, formulated as:

$$\begin{aligned}
\mu_{ij} &= \mathbf{W}_\mu^\intercal(\mathbf{W}_A^\intercal A_i + \mathbf{W}_M^\intercal M_{ij}),\\
\gamma_{ij} &= \frac{\exp(\mu_{ij})}{\sum_{t=1}^{T}\exp(\mu_{it})},\\
D_i &= \sum\nolimits_{j=1}^{T}\gamma_{ij}M_{ij},
\end{aligned} \tag{2}$$

where $\mathbf{W}_A^\intercal$, $\mathbf{W}_M^\intercal$ and $\mathbf{W}_\mu^\intercal$ are projection matrices.

Intra-Modal Attention. As mentioned in Section 3.1 and 3.2, our method predicts the boundary confidence of the current frame I_t based on a clip U centered on it. In the clip, features of different timestamps should not be equally important. For example, the center frame of the clip is more important than edge frames of the clip in most cases. To adaptively aggregate and enhance key representations of RGB features A and DDM features D, we employ two sets of ω *learnable* queries $\mathbf{q}$, which are formed by adding content queries $\mathbf{c}_q$ (initialized with standard normal distribution) and learnable positional embeddings of queries $\mathbf{p}_q$. Specifically, We exploit two separate transformer decoders to respectively aggregate and enhance key intra-modal features of A and D,

$$\begin{aligned}
&\mathbf{q} = \mathbf{c}_q + \mathbf{p}_q,\\
&\mathbf{k} = \mathbf{c}_k + \mathbf{p}_k = H + \mathbf{p}_k, \quad \mathbf{v} = \mathbf{c}_v = H,
\end{aligned} \tag{3}$$

where $\mathbf{c}_k$ and $\mathbf{c}_v$ are features H of the modality (A or D), $\mathbf{p}_k$ is sine positional embedding. In cross-attention layers, queries globally attend and aggregate features of high activation into each query. Self-attention layers model the dependencies between queries and enhance corresponding query embeddings. Through intra-modal representation learning, two sets of queries $\mathbf{q}$ independently aggregate and enhance key features of two modalities, and become refined queries $\mathbf{q}^{'}$. In Table 4e, we observe that cross-modal attention can achieve better performance upon refined key features $\mathbf{q}^{'}$, compared with the unrefined features H.

Cross-Modal Attention. Due to the diversity and complex semantics of generic event boundaries, it is difficult to distinguish them with only appearance or motion features. A fusion of them can alleviate this issue, but previous fusion methods (*e.g.*, feature concatenation) fail to jointly learn features across modalities and make full use of feature complementarity. Thus, in order to leverage the dependencies between two modalities, we perform cross-modal feature aggregation. Concretely, we take the *feature pair* of ω refined queries $\mathbf{q}^{'}$ as the input of two independent co-attention transformers. One co-attention transformer takes refined RGB features $\mathbf{q}_A^{'}$ as queries, and refined DDM features $\mathbf{q}_D^{'}$ as keys and values,

$$\begin{aligned}
&\mathbf{q} = \mathbf{c}_q = \mathbf{q}_A^{'},\\
&\mathbf{k} = \mathbf{c}_k = \mathbf{q}_D^{'}, \quad \mathbf{v} = \mathbf{c}_v = \mathbf{q}_D^{'}.
\end{aligned} \tag{4}$$

That is to say, $\mathbf{q}_A^{'}$ guide and enhance $\mathbf{q}_D^{'}$ via cross-attention layers. Inputs of the other co-attention transformer are symmetric to the first one, namely $\mathbf{q}_D^{'}$ as queries, $\mathbf{q}_A^{'}$ as keys and values. Through cross-attention layers, cross-modal attention module outputs RGB-conditioned DDM features $\mathbf{q}_D^{''}$ and DDM-modulated RGB features $\mathbf{q}_A^{''}$. As a consequence, DDM-Net aggregates appearance and motion cues with cross-modal guidance, effectively improving the discrimination of event boundaries.

3.5. Training

Balanced Sampler. GEBD is a binary classification task, and non-boundary frames are far more than boundary frames (r:1). Following [35], we leverage a balanced sampler. Due to slowness prior in videos [50], features of consecutive non-boundary frames change at a very slow speed. Hence, we apply a sparse sampling strategy on non-boundary frames, namely select one out of sequential r non-boundary frames randomly and sample all boundary frames.

Rel.Dis. threshold		0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	avg
Unsuper.	SceneDetect [5]	0.275	0.300	0.312	0.319	0.324	0.327	0.330	0.332	0.334	0.335	0.318
	PA - Random [35]	0.336	0.435	0.484	0.512	0.529	0.541	0.548	0.554	0.558	0.561	0.506
	PA [35]	0.396	0.488	0.520	0.534	0.544	0.550	0.555	0.558	0.561	0.564	0.527
Super.	BMN [24]	0.186	0.204	0.213	0.220	0.226	0.230	0.233	0.237	0.239	0.241	0.223
	BMN-StartEnd [24]	0.491	0.589	0.627	0.648	0.660	0.668	0.674	0.678	0.681	0.683	0.640
	TCN-TAPOS [20]	0.464	0.560	0.602	0.628	0.645	0.659	0.669	0.676	0.682	0.687	0.627
	TCN [20]	0.588	0.657	0.679	0.691	0.698	0.703	0.706	0.708	0.710	0.712	0.685
	PC [35]	0.625	0.758	0.804	0.829	0.844	0.853	0.859	0.864	0.867	0.870	0.817
	DDM-Net	**0.764**	**0.843**	**0.866**	**0.880**	**0.887**	**0.892**	**0.895**	**0.898**	**0.900**	**0.902**	**0.873**

Table 1. Comparison with previous methods on the validation set of Kinetics-GEBD, measured by F1 score at different Rel.Dis. thresholds.

Rel.Dis. threshold		0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	avg
Unsuper.	SceneDetect [5]	0.035	0.045	0.047	0.051	0.053	0.054	0.055	0.056	0.057	0.058	0.051
	PA - Random [35]	0.158	0.233	0.273	0.310	0.331	0.347	0.357	0.369	0.376	0.384	0.314
	PA [35]	0.360	0.459	0.507	0.543	0.567	0.579	0.592	0.601	0.609	0.615	0.543
Super.	ISBA [7]	0.106	0.170	0.227	0.265	0.298	0.326	0.348	0.369	0.382	0.396	0.302
	TCN [20]	0.237	0.312	0.331	0.339	0.342	0.344	0.347	0.348	0.348	0.348	0.330
	CTM [16]	0.244	0.312	0.336	0.351	0.361	0.369	0.374	0.381	0.383	0.385	0.350
	TransParser [34]	0.289	0.381	0.435	0.475	0.500	0.514	0.527	0.534	0.540	0.545	0.474
	PC [35]	0.522	0.595	0.628	0.646	0.659	0.665	0.671	0.676	0.679	0.683	0.642
	DDM-Net	**0.604**	**0.681**	**0.715**	**0.735**	**0.747**	**0.753**	**0.757**	**0.760**	**0.763**	**0.767**	**0.728**

Table 2. Comparison with previous GEBD methods on the validation set of TAPOS, measured by F1 score at different Rel.Dis. thresholds.

Method	rank1 [19]	rank2 [15]	rank3 [38]	DDM-Net
F1 score	0.8354	0.8330	0.8309	**0.8368**

Table 3. Comparison with winner solutions of LOVEU challenge on the testing set of Kinetics-GEBD, measured by F1 score@0.05.

Loss Function. The balanced sampler relieves the imbalance of positive and negative samples. Therefore, we simply define the binary classification loss $\mathcal{L}_{bc}$ as:

$$\mathcal{L}_{bc} = -\frac{1}{N}\sum_{\eta=1}^{N}(\hat{p}_\eta \log p_\eta + (1-\hat{p}_\eta)\log(1-p_\eta)), \quad (5)$$

where p_η is the binary classification probability and N is the total number of training samples. $\hat{p}_\eta$ is 1 if the sample is marked as a boundary, and otherwise 0.

3.6. Inference

Linear Fusion of Logits. After progressive attention, $\mathbf{q}_A^{''}$ and $\mathbf{q}_D^{''}$ are separately passed into two independent fc layers to generate logits l_A and l_D. With a learnable parameter α, we perform a linear fusion of logits: $l = \alpha * l_A + (1-\alpha) * l_D$. Softmax function is applied on the final logit l to get the boundary probability p.

Efficient Post-processing Scheme. Repeating the above process of predicting the boundary probability of one frame, we obtain the boundary confidence sequence of the whole video. To select the final boundary predictions of the video, we apply an efficient post-processing scheme on the sequence. In detail, a boundary frame should satisfy the following two requirements: (1) The boundary probability of the frame is greater than a set threshold θ (*e.g.*, 0.5). (2) Its boundary probability is the maximum within a pre-defined range (*e.g.*, [-5, 5]). Since our post-processing scheme is free of time-consuming pairwise IoU calculation, it only takes about 0.0003 seconds per video (5.302s for all 18,813 videos) on one Nvidia V100 machine.

4. Experiments

4.1. Dataset and Setup

Kinetics-GEBD. Kinetics-GEBD dataset [35] consists of 60,000 videos randomly selected from Kinetics-400. Among them, 18,794 training videos and 17,725 testing videos are randomly selected from Kinetics-400 training set. Kinetics-GEBD validation set contains all 18,813 videos in Kinetics-400 validation set. The ratio of training, validation and testing sets is nearly 1:1:1. Since temporal annotations of testing set are not available, we train on the training set and evaluate with the validation set.

TAPOS. TAPOS dataset [34] contains Olympics sports videos with 21 actions. There are 13,094 training action instances and 1,790 validation action instances. Following [35], we re-purpose TAPOS for GEBD task by trimming each action instance with its action label hidden.

Evaluation Protocol. To evaluate the results of generic event boundary detection task, we calculate F1 score and

Representation	0.05	0.25	0.5	Average
RGB	0.6793	0.8589	0.8772	0.8375
Optical flow	0.6625	0.8045	0.8206	0.7877
RGB differences	0.7272	0.8591	0.8753	0.8440
DDM	0.7512	0.8738	0.8861	0.8591
RGB + Optical flow (two-stream)	0.6881	0.8682	0.8844	0.8465
RGB + RGB differences (two-stream)	0.7307	0.8702	0.8834	0.8536
RGB + DDM (on-the-fly)	**0.7643**	**0.8870**	**0.9016**	**0.8726**

(a) **Study on Different Representations.** We compare the performance of single modality and two modalities. When combined with RGB features, only DDM is calculated online and can be trained on-the-fly.

w	s	0.05	0.25	0.5	Average
5	3	0.7551	0.8815	0.8970	0.8666
5	6	**0.7643**	**0.8870**	**0.9016**	**0.8726**
5	12	0.7521	0.8788	0.8924	0.8636
5	30	0.6914	0.8563	0.8717	0.8371
15	2	**0.7703**	**0.8890**	**0.9042**	**0.8754**

(b) **Study on Temporal Views.** We compare the performance of input clips with different temporal views, and further boost the performance with a more-frames setting.

Multi-level	0.05	0.25	0.5	Average
None	0.7353	0.8617	0.8726	0.8463
Spatial	0.7487	0.8694	0.8820	0.8552
Temporal	0.7497	0.8727	0.8848	0.8579
Spatial + Temporal	**0.7643**	**0.8870**	**0.9016**	**0.8726**

(c) **Study on Multi-Level Feature Bank.** 'None' refers to only using the feature collected from the last layer of backbone network (layer4 of ResNet50).

Operators	0.05	0.25	0.5	Average
Manhattan	0.7632	0.8870	**0.9024**	0.8725
Euclidean	**0.7643**	**0.8870**	0.9016	**0.8726**
Chebyshev	0.7540	0.8789	0.8931	0.8643
Cosine	0.7611	0.8834	0.8982	0.8691

(d) **Study on Difference Operators.** 'Difference operator' refers to the distance metric applied in difference calculation.

Aggregation	0.05	0.25	0.5	Average
Avg	0.7498	0.8685	0.8804	0.8543
Intra	0.7588	0.8793	0.8922	0.8650
Cross	0.7590	0.8770	0.8894	0.8631
Intra + Cross	**0.7643**	**0.8870**	**0.9016**	**0.8726**

(e) **Study on Aggregation Methods of Progressive Attention**. 'Avg' refers to direct temporal averaging operation without attention.

Table 4. **Ablation Studies on Kinetics-GEBD dataset,** measured by F1 score at different Rel.Dis. thresholds.

use the Relative Distance (Rel.Dis.) measurement [35]. Rel.Dis. is the relative distance between predictions and ground truths, divided by the length of the corresponding video. Given a threshold, a prediction is determined to be true if Rel.Dis. is smaller than or equal to the threshold, otherwise false. In experiments, we follow [35] to report F1 score with Rel.Dis. threshold set [0.05 : 0.05 : 0.5].

Implementation Details. In practice, we select one frame out of every 3 consecutive frames, namely the stride of boundary evaluation is 3. To predict the confidence of the current frame, we take a $T \times s$ ($T = 2 \times w + 1$) clip as the input, where w is 5 and s is 6. Following [35], our model is built on ImageNet-pretrained ResNet-50 backbone and trained *end-to-end*. m and n of multi-level feature bank are set to 3. ω of progressive attention is set to 5. To train DDM-Net, we employ Adam as the optimizer. The batch size is set to 32 and the learning rate is set to 1e-5.

4.2. Main Results

We fairly compare our DDM-Net with state-of-the-art methods on the validation set of Kinetics-GEBD and TAPOS. As a result, our method outperforms the state-of-the-art methods by a large margin at all Rel.Dis. thresholds.

Kinetics-GEBD. Table 1 illustrates the performance of different methods on Kinetics-GEBD validation set. It can be seen that DDM-Net remarkably outperforms other methods on F1 score, demonstrating the effectiveness of dense differences and sophisticated aggregation. Especially, DDM-Net achieves significant improvements from 62.5% to 76.4% at the most strict threshold (Rel.Dis.=0.05). A boost of nearly 14 percent proves the boundary predictions of DDM-Net are the most precise. Furthermore, combined with a more powerful backbone CSN [42], our DDM-Net can be superior to winner solutions of LOVEU Challenge@CVPR 2021, as shown in Table 3. It is worth noting that this result is obtained without bells and whistles (*e.g.*, model ensemble, audio data and human-object detector) of winner solutions.

TAPOS. The comparison results of the state-of-the-art GEBD methods on TAPOS are summarized in Table 2. Since DDM-Net is able to learn complex semantics and distinguish subtle changes between sub-actions, it obtains state-of-the-art performance on TAPOS, increasing F1 score@0.05 from 52.2% to 60.4%. The result proves that our model can not only achieve accurate generic event boundary detection (Kinetics-GEBD), but also precisely detect boundaries between fine-grained sub-actions (TAPOS).

4.3. Ablation Study

Study on Different Representations. We analyze our proposed DDM by experimenting with different representations, which is shown in Table 4a. *First*, we compare the performance of single representation. Our proposed DDM outperforms RGB, optical flow and RGB differences by a large margin, especially under strict settings (Rel.Dis. = 0.05). *Second*, we find that DDM brings the greatest improvements when combined with RGB features, which proves the complementarity of dense differences and RGB features. Furthermore, as DDM is calculated on-the-fly, our method is free of training two separate networks and thus is more efficient than previous two-stream methods.

Study on Temporal Views. In Section 3.2, we argue that boundary frames have stronger correlations with their adjacent snippets than faraway frames. To prove our point, we experiment with clips of different temporal views. As illustrated in Table 4b, the temporal view grows with the increase of stride s, but the performance decreases in row 3

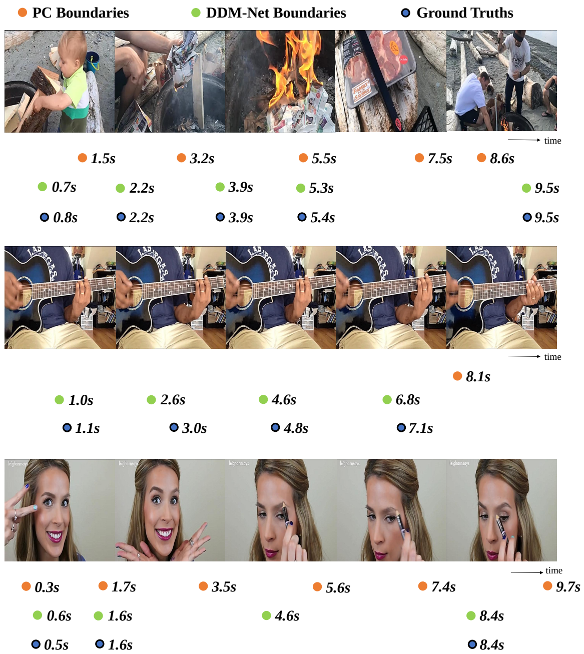

Figure 3. **Qualitative results** and comparisons of PC, DDM-Net and ground truths on Kinetics-GEBD dataset.

and 4. Particularly, when s is increased to 30, namely we classify the current frame with uniformly sampled frames through the whole video (about 300 frames), the performance witnesses a significant drop. Hence, it is unnecessary to introduce global temporal contexts in GEBD task. In addition, if the temporal view is fixed, using more frames ($w = 15$ instead of $w = 5$) to calculate denser differences can slightly improve the performance (row 2 and row 5).

Study on Multi-Level Feature Bank. In order to model various motion patterns of diverse event boundaries, we first build a multi-level feature bank, then calculate dense differences upon it. As demonstrated in Table 4c, both multi-level spatial and temporal features outperform the result with single-level features. Moreover, a combination of multi-level spatial and temporal features (firstly generate multi-level spatial features, then yield multi-level temporal features upon spatial features of each layer) can further lead to a remarkable performance advance, confirming the overall contributions of multi-level feature bank.

Study on Difference Operators. In Table 4d, Euclidean distance, Manhattan distance and Cosine distance are respectively employed to measure the feature difference. We observe that our method obtains close performance. Therefore, performance of DDM-Net is robust to the choice of difference operators. In contrast, our model witnesses a drop with Chebyshev distance. Since Chebyshev distance measures the max discrepancy among all channels, it cannot reflect the overall differences between two frames.

Study on Aggregation Methods of Progressive Attention. In Table 4e, we study two aggregation methods of progressive attention (map-squeezed attention is required to align the shape and cannot be removed). Intra-modal attention mainly focuses on aggregating and enhancing key intra-modal features with learnable queries $\mathbf{q}$. Compared with the overall feature sequence H, refined queries $\mathbf{q}^{'}$ contain key patterns and less noise, which can explain the performance gain of row 2. If only cross-modal attention is leveraged, overall RGB features A and DDM features D directly query and guide each other through joint feature learning across modalities, leading to the improvement of row 3. Moreover, instead of H, cross-modal aggregation of refined key features $\mathbf{q}^{'}$ can further boost the performance (row 4).

4.4. Qualitative Results

Figure 3 displays qualitative results of our method, including different types of event boundaries. The first example is a video of several shot changes. DDM-Net precisely perceives temporal variations and hits every boundary instance, while predictions of PC are not accurate. Case in the second row is more challenging, as only the position of the left hand changes. DDM-Net is able to model complex semantics and distinguish subtle action changes, therefore it makes accurate predictions. In contrast, PC misses all the ground truths. The last example is a combination of shot changes and action changes. Since our method is more robust to temporal noise (camera jitter), it predicts fewer false positives. In summary, thanks to multi-level DDM and progressive attention, our method is able to precisely perceive temporal changes and understand complicated semantics, hence it has shown advantages in many different cases.

5. Conclusion

In this paper, we have presented a modular framework for the task of generic event boundary detection (GEBD). To perceive diverse temporal variations and learn complex semantics of generic event boundaries, our method progressively attends to multi-level dense difference maps (DDM). Thanks to holistic temporal modeling and joint feature learning across modalities, our DDM-Net outperforms the previous state-of-the-art methods by a large margin on Kinetics-GEBD and TAPOS benchmark. In addition, our method is better than winner solutions of LOVEU Challenge@CVPR 2021, further demonstrating the efficacy of DDM-Net. As for **limitations**, large-scale GEBD benchmarks of untrimmed videos are expected to further validate our method in future work.

Acknowledgements. This work is supported by National Natural Science Foundation of China (No.62076119, No.61921006), Program for Innovative Talents and Entrepreneur in Jiangsu Province, and Collaborative Innovation Center of Novel Software Technology and Industrialization.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, pages 6077–6086. Computer Vision Foundation / IEEE Computer Society, 2018. 3

[2] Lorenzo Baraldi, Costantino Grana, and Rita Cucchiara. Shot and scene detection via hierarchical clustering for re-using broadcast video. In *CAIP (1)*, volume 9256 of *Lecture Notes in Computer Science*, pages 801–811. Springer, 2015. 2

[3] Shyamal Buch, Victor Escorcia, Bernard Ghanem, Li Fei-Fei, and Juan Carlos Niebles. End-to-end, single-stream temporal action detection in untrimmed videos. In *BMVC*. BMVA Press, 2017. 2

[4] João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *CVPR*, pages 4724–4733. IEEE Computer Society, 2017. 2

[5] Brandon Castellano. PySceneDetect: an intelligent scene cut detection and video splitting tool. `https://github.com/Breakthrough/PySceneDetect`, 2014. 6

[6] Yi-Wen Chen, Yi-Hsuan Tsai, and Ming-Hsuan Yang. End-to-end multi-modal video temporal grounding. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. 3

[7] Li Ding and Chenliang Xu. Weakly-supervised action segmentation with iterative soft boundary assignment. In *CVPR*, pages 6508–6516. Computer Vision Foundation / IEEE Computer Society, 2018. 6

[8] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Counting out time: Class agnostic video repetition counting in the wild. In *CVPR*, pages 10384–10393. Computer Vision Foundation / IEEE, 2020. 3

[9] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *CVPR*, pages 1933–1941. IEEE Computer Society, 2016. 3, 4

[10] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. TALL: temporal activity localization via language query. In *ICCV*, pages 5277–5285. IEEE Computer Society, 2017. 3

[11] Ruohan Gao, Rogério Schmidt Feris, and Kristen Grauman. Learning to separate object sounds by watching unlabeled video. In *ECCV (3)*, volume 11207 of *Lecture Notes in Computer Science*, pages 36–54. Springer, 2018. 3

[12] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *ICCV*, pages 1705–1714. IEEE, 2019. 2

[13] Michael Gygli. Ridiculously fast shot boundary detection with fully convolutional neural networks. In *CBMI*, pages 1–4. IEEE, 2018. 2

[14] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan C. Russell. Localizing moments in video with natural language. In *ICCV*, pages 5804–5813. IEEE Computer Society, 2017. 3

[15] Dexiang Hong, Congcong Li, Longyin Wen, Xinyao Wang, and Libo Zhang. Generic event boundary detection challenge at cvpr 2021 technical report: Cascaded temporal attention network (castanet), 2021. 6

[16] De-An Huang, Li Fei-Fei, and Juan Carlos Niebles. Connectionist temporal modeling for weakly supervised action labeling. In *ECCV (4)*, volume 9908 of *Lecture Notes in Computer Science*, pages 137–153. Springer, 2016. 6

[17] Boyuan Jiang, Mengmeng Wang, Weihao Gan, Wei Wu, and Junjie Yan. STM: spatiotemporal and motion encoding for action recognition. In *ICCV*, pages 2000–2009. IEEE, 2019. 3

[18] Imran N. Junejo, Emilie Dexter, Ivan Laptev, and Patrick Pérez. View-independent action recognition from temporal self-similarities. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(1):172–185, 2011. 3

[19] Hyolim Kang, Jinwoo Kim, Kyungmin Kim, Taehyun Kim, and Seon Joo Kim. Winning the cvpr'2021 kinetics-gebd challenge: Contrastive learning approach, 2021. 6

[20] Colin Lea, Austin Reiter, René Vidal, and Gregory D. Hager. Segmental spatiotemporal cnns for fine-grained action segmentation. In *ECCV (3)*, volume 9907 of *Lecture Notes in Computer Science*, pages 36–52. Springer, 2016. 6

[21] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L. Berg. TVQA: localized, compositional video question answering. In *EMNLP*, pages 1369–1379. Association for Computational Linguistics, 2018. 3

[22] Yan Li, Bin Ji, Xintian Shi, Jianguo Zhang, Bin Kang, and Limin Wang. TEA: temporal excitation and aggregation for action recognition. In *CVPR*, pages 906–915. Computer Vision Foundation / IEEE, 2020. 3

[23] Chuming Lin, Chengming Xu, Donghao Luo, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yanwei Fu. Learning salient boundary feature for anchor-free temporal action localization. In *CVPR*, pages 3320–3329. Computer Vision Foundation / IEEE, 2021. 2

[24] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. BMN: boundary-matching network for temporal action proposal generation. In *ICCV*, pages 3888–3897. IEEE, 2019. 2, 6

[25] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. BSN: boundary sensitive network for temporal action proposal generation. In *ECCV (4)*, volume 11208 of *Lecture Notes in Computer Science*, pages 3–21. Springer, 2018. 2

[26] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection - A new baseline. In *CVPR*, pages 6536–6545. Computer Vision Foundation / IEEE Computer Society, 2018. 2

[27] Xiaolong Liu, Yao Hu, Song Bai, Fei Ding, Xiang Bai, and Philip H. S. Torr. Multi-shot temporal event localization: A benchmark. In *CVPR*, pages 12596–12606. Computer Vision Foundation / IEEE, 2021. 2

[28] Zhaoyang Liu, Donghao Luo, Yabiao Wang, Limin Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Tong Lu. Teinet: Towards an efficient architecture for video recognition. In *AAAI*, pages 11669–11676. AAAI Press, 2020. 3

[29] Zhaoyang Liu, Limin Wang, Wayne Wu, Chen Qian, and Tong Lu. Tam: Temporal adaptive module for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13708–13718, October 2021. 3

[30] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, pages 13–23, 2019. 3

[31] Yiwei Lu, Frank Yu, Mahesh Kumar Krishna Reddy, and Yang Wang. Few-shot scene-adaptive anomaly detection. In *ECCV (5)*, volume 12350 of *Lecture Notes in Computer Science*, pages 125–141. Springer, 2020. 2

[32] Chenxu Luo and Alan L. Yuille. Grouped spatial-temporal aggregation for efficient action recognition. In *ICCV*, pages 5511–5520. IEEE, 2019. 3

[33] Zhiwu Qing, Haisheng Su, Weihao Gan, Dongliang Wang, Wei Wu, Xiang Wang, Yu Qiao, Junjie Yan, Changxin Gao, and Nong Sang. Temporal context aggregation network for temporal action proposal refinement. In *CVPR*, pages 485–494. Computer Vision Foundation / IEEE, 2021. 2

[34] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Intra- and inter-action understanding via temporal action parsing. In *CVPR*, pages 727–736. Computer Vision Foundation / IEEE, 2020. 2, 6

[35] Mike Zheng Shou, Stan Weixian Lei, Weiyao Wang, Deepti Ghadiyaram, and Matt Feiszli. Generic event boundary detection: A benchmark for event segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8075–8084, October 2021. 1, 2, 3, 5, 6, 7

[36] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, pages 568–576, 2014. 2, 3, 4

[37] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *CVPR*, pages 6479–6488. Computer Vision Foundation / IEEE Computer Society, 2018. 2

[38] Quan Sun, Shimin Chen, Chen Chen, Xunqiang Tao, and Yandong Guo. Generic event boundary detection: Submission to loveu challenge 2021. https://github.com/VisualAnalysisOfHumans/LOVEU_TRACK1_TOP3_SUBMISSION, 2021. 6

[39] Jing Tan, Jiaqi Tang, Limin Wang, and Gangshan Wu. Relaxed transformer decoders for direct action proposal generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13526–13535, October 2021. 2

[40] Shitao Tang, Litong Feng, Zhanghui Kuang, Yimin Chen, and Wei Zhang. Fast video shot transition localization with deep structured models. In *ACCV (1)*, volume 11361 of *Lecture Notes in Computer Science*, pages 577–592. Springer, 2018. 2

[41] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *ECCV (2)*, volume 11206 of *Lecture Notes in Computer Science*, pages 252–268. Springer, 2018. 3

[42] Du Tran, Heng Wang, Matt Feiszli, and Lorenzo Torresani. Video classification with channel-separated convolutional networks. In *ICCV*, pages 5551–5560. IEEE, 2019. 7

[43] Barbara Tversky and Jeffrey M Zacks. Event perception. *Oxford handbook of cognitive psychology*, 2013. 1

[44] Limin Wang, Zhan Tong, Bin Ji, and Gangshan Wu. TDN: temporal difference networks for efficient action recognition. In *CVPR*, pages 1895–1904. Computer Vision Foundation / IEEE, 2021. 3

[45] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV (8)*, volume 9912 of *Lecture Notes in Computer Science*, pages 20–36. Springer, 2016. 2, 3, 4

[46] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 2048–2057. JMLR.org, 2015. 3

[47] Mengmeng Xu, Chen Zhao, David S. Rojas, Ali K. Thabet, and Bernard Ghanem. G-TAD: sub-graph localization for temporal action detection. In *CVPR*, pages 10153–10162. Computer Vision Foundation / IEEE, 2020. 2

[48] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alexander J. Smola. Stacked attention networks for image question answering. In *CVPR*, pages 21–29. IEEE Computer Society, 2016. 3

[49] Runhao Zeng, Wenbing Huang, Chuang Gan, Mingkui Tan, Yu Rong, Peilin Zhao, and Junzhou Huang. Graph convolutional networks for temporal action localization. In *ICCV*, pages 7093–7102. IEEE, 2019. 2

[50] Zhang Zhang and Dacheng Tao. Slow feature analysis for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(3):436–450, 2012. 5

[51] Yue Zhao, Yuanjun Xiong, and Dahua Lin. Recognize actions by disentangling components of dynamics. In *CVPR*, pages 6566–6575. Computer Vision Foundation / IEEE Computer Society, 2018. 3

[52] Yuan Zhi, Zhan Tong, Limin Wang, and Gangshan Wu. Mgsampler: An explainable sampling strategy for video action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1513–1522, October 2021. 3