NewsNet: A Novel Dataset for Hierarchical Temporal Segmentation

Haoqian Wu[1†], Keyu Chen[1†], Haozhe Liu[2†], Mingchen Zhuge[2†], Bing Li[2✉],
Ruizhi Qiao[1✉], Xiujun Shu[1], Bei Gan[1], Liangsheng Xu[1], Bo Ren[1], Mengmeng Xu[2], Wentian Zhang[2]
Raghavendra Ramachandra[3], Chia-Wen Lin[4], Bernard Ghanem[2]

[1] Tencent [2] AI Initiative, King Abdullah University of Science and Technology (KAUST)
[3] Norwegian University of Science and Technology (NTNU)
[4] National Tsing Hua University (NTHU)

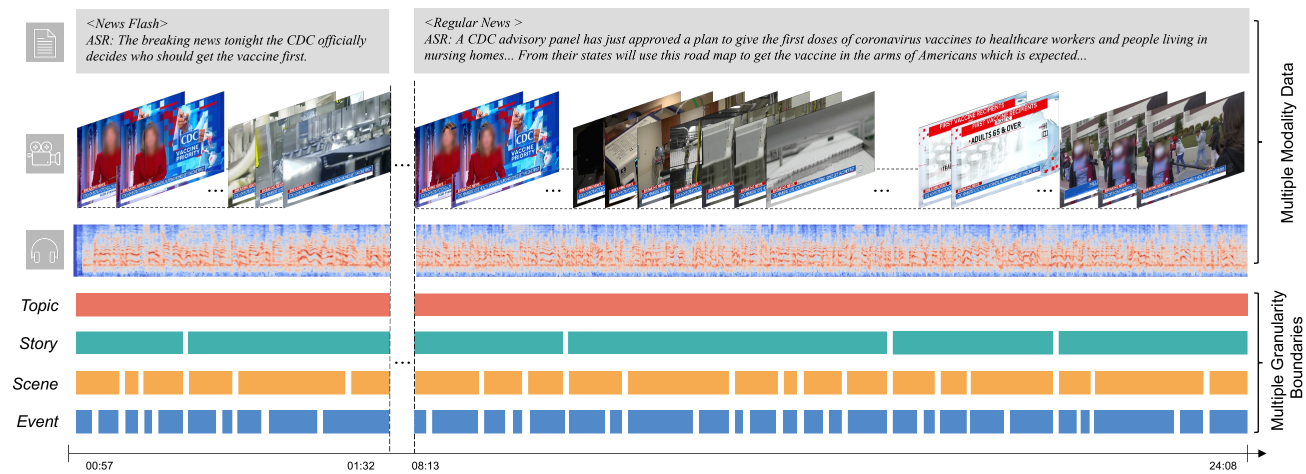

Figure 1. **Examples from the proposed NewsNet.** The dataset provides multimodal and hierarchical labels on each video, including 4-level hierarchical temporal information (*i.e.*, event, scene, story, topic) and 3 modality-specific cues (*i.e.*, video, audio, text). A prerequisite motivation behind NewsNet is that long-form videos can be recursively represented into several sub-videos according to the semantics, which is an intrinsic property of the video but rarely explored in the community. To bridge this research gap, this paper innovatively constructs a large-scale dataset to benchmark the algorithms for hierarchical understanding like a human being.

Abstract

Temporal video segmentation is the get-to-go automatic video analysis, which decomposes a long-form video into smaller components for the following-up understanding tasks. Recent works have studied several levels of granularity to segment a video, such as shot, event, and scene. Those segmentations can help compare the semantics in the corresponding scales, but lack a wider view of larger temporal spans, especially when the video is complex and structured. Therefore, we present two abstractive levels of temporal segmentations and study their hierarchy to the existing fine-grained levels. Accordingly, we collect NewsNet, the largest news video dataset consisting of 1,000 videos in over 900 hours, associated with several tasks for hierarchical temporal video segmentation. Each news video is a collection of stories on different topics, represented as aligned audio, visual, and textual data, along with extensive frame-wise annotations in four granularities. We assert that the study on NewsNet can advance the understanding of complex structured video and benefit more areas such as short-video creation, personalized advertisement, digital instruction, and education. Our dataset and code is publicly available at https://github.com/NewsNet-Benchmark/NewsNet.

1. Introduction

Temporal video segmentation is a critical problem in video understanding, which is essential for many video applications such as video classification [10, 15, 37, 58, 59, 62],

† Equal Contribution
Corresponding Authors: Bing Li (bing.li@kaust.edu.sa) and Ruizhi Qiao (ruizhiqiao@tencent.com).

Table 1. Data comparison between the NewsNet and other related datasets. The NewsNet provides various multimodal data and hierarchical temporal segmentation annotations. Doc: documentary, Ads: advertisement. (*Please refer to our project page for more details.*)

Dataset	# Video	Duration (hours)	Modality	# Annotation(s) per Video				Source
				Topic	Story	Scene	Event	
AVS [75]	197	-	Visual	-	-	-	14.2	Ads
BBC [6]	11	9	Visual	-	-	49.7	-	Doc
OVSD [48]	21	10	Visual	-	-	28.9	-	Generic
Kinetics-GEBD [51]	54,691	152	Audio + Visual	-	-	-	4.9	Action
MovieNet [23] †	1,100	2174	Text + Audio + Visual	-	-	66.0	849.1	Movie
RAI [7]	10	-	Visual	-	-	-	98.7	News
TI-News [35]	477	244	Audio + Visual	-	55.6	-	530.4	News
NewsNet (Ours)	1,000	946	Text + Audio + Visual	8.5	51.6	87.9	654.4	News

† The number of annotations for *Scene* and *Shot* is counted from MovieScene [47], which is a subset of MovieNet.

captioning [3, 34, 63, 71] and retrieval [5, 16, 31, 52]. Temporal video segmentation aims to group successive video frames into short segments along the temporal dimension. With the explosive growth of long-form videos, it is desirable that temporal video segmentation can convert a video into more meaningful segments for more efficient access to the video. However, it is challenging to develop effective temporal segmentation tools for long-form videos, since this requires a comprehensive understanding of video structure, while long-form videos contain complex content.

Towards temporal video segmentation, existing works explore shot, event, and scene segmentation tasks, respectively. Shot segmentation [38, 55, 57] divides a video into shots, where a shot consists of consecutive and visually continuous frames captured by a camera without interruption [23]. Yet, shot segmentation only considers low-level visual cues (*i.e*. visual similarity), lacking semantic understanding. Instead, event segmentation [25, 51, 56] divides a video by detecting the moments of changes such as action/subject changes. To better capture the underlying structure of a video, recent works [12, 47, 65] introduce video scene segmentation which segments a video into scene segments, each comprising successive shots semantically related to the same scene. Scene segmentation enables a coarser and higher-level representation than shot segmentation. However, compared to the rich content of massive long-form videos, scene/event is fine-grained and often lacks a high-level summarization of video content, which is insufficient for capturing the complex semantic structure of many videos and briefly representing video content.

In this work, we first explore how to comprehensively represent the complex structure of a long-form video for temporal video segmentation. Humans can hierarchically divide a video into segments of different granularities according to multi-level semantic information (*e.g.* scene and topic), from the perspective of cognitive science. Natural language processing researchers have widely explored topic-level understanding [29, 41] for summarizing documents [8, 42], while little effort has been devoted to long-form videos. Inspired by these observations, besides scene and event, we propose to introduce two higher-level semantics (*i.e*. story and topic) into temporal video segmentation, to provide a brief and semantic structure representation. As a result, such hierarchical and multi-level understanding brings about scalable video structure representation for temporal video segmentation on long-form videos. That is, a long-form video can be split into finer segments with lower-level semantics (*e.g*. scene), but also can be summarized into coarser ones yet with higher-level semantics (*e.g*. topic) by recursively grouping finer segments, which comprehensively represents video structures from coarse to fine.

However, the community lacks high-quality datasets to conduct this research. In particular, as shown in Table 1, most datasets only provide temporal structure annotations with regard to events or scenes. TI-News [35] and MovieScene [47] provide two levels of annotations, but these datasets lack topic-level ones.

To effectively break this limitation, we build a novel large-scale dataset for hierarchical temporal segmentation, named NewsNet. The unique properties of our NewsNet introduce many advantages. First, it is among the largest datasets in the news domain. We collect over 900 hours of videos from 20 mainstream news platforms. It has a highly diverse distribution of data. Second, we carefully annotated it frame-by-frame with 4 hierarchical levels to ensure its quality can meet our needs. Third, it is multimodal, including textual, visual, and audio information. Due to the nature of the news, the alignment across modalities is accurate, which makes multimodal joint training of models feasible. Finally, the videos in NewsNet provide a complete understanding of public events. Compared with other video datasets [4, 23, 26], it introduces more objective open-world knowledge (*e.g.*, news introduction) while including subjective factual commentary (*e.g.*, host comments on news events), making it more amenable to real-life application.

Based on NewsNet, we empirically highlight two promising directions for long-span temporal segmentation: 1) Infusing Multi-Modality knowledge can significantly improve the performance of long-form temporal segmentation; 2) Although story- and topic-level segmentation is challenging, it can be benefited from hierarchical modeling with the event- and scene-level segmentation tasks.

The main contributions of this paper are as follows:

- We propose a novel large-scale dataset NewsNet for long-form video structure understanding. This dataset is derived from 900+ hours of video and annotated with 4 hierarchical levels of semantics.
- NewsNet provides dense annotations and multi-modal information, promoting diverse benchmarks: separate/hierarchical temporal video segmentation in scene/story/topic levels, as well as other common tasks like classification, video localization/grounding, and highlight detection.
- We formulate a new benchmark, *i.e.,* hierarchical modeling in the temporal segmentation task, which needs a single model to predict segments of multiple hierarchical levels. Based on the empirical study, we bring insights into how hierarchical modeling potentially benefits the temporal video segmentation task, which was almost never discussed.

2. Related Work

2.1. Related Dataset and Benchmark

To better introduce our work toward hierarchical temporal segmentation, we first review research related to temporal segmentation and video challenges. As shown in Table 1, extensive benchmarks have been proposed to test the performance of temporal video segmentation. More recently, Shou et al. [51] argues that humans can capture the boundary without predefined target classes, and thus proposed Generic Event Boundary Detection (GEBD) as a new benchmark for detecting generic event boundaries on Kinetics-400 [26]. A few works [25, 33, 43–45, 56] have been proposed to attempt to address this challenge, which is a very active sub-field in long-form video understanding [1, 2]. In this paper, as the NewsNet densely annotates all temporal boundaries, including events like GEBD, scenes like MovieScene [47], and stories like TI-News [35]. NewsNet can test the performance of the existing tasks on almost all temporal localization and segmentation tasks. Moreover, with such hierarchical labels, we can further investigate hierarchical modeling in long-form video understanding, which is an important but rarely explored field.

As an overview of the existing benchmarks in NLP and CV, the temporal annotations are labeled in one or two levels, only focusing on the global abstraction or local understanding. However, the video is naturally hierarchical, which can be recursively represented into sub-units. Therefore, the topic-, story-, scene- and event-level tasks should not be drawn into individual lines, but work in a harmonious way. From such a motivation, we contribute a new dataset called NewsNet with hierarchical dense temporal annotations to the community. Since the annotation cost for different groups is quite different, the scale for topic-level benchmarks [4, 9, 23, 26] is generally larger than that of the other datasets [6, 7, 35, 47, 48, 75]. This paper tries our best to reach a satisfactory trade-off between scale and the quality of the annotations. The proposed dataset does not only cover 1000 videos with 946 hours but also achieves 654.4 event-level annotations per video, which is very competitive over the existing baselines. Moreover, to further generalize the promotion, textual, audio, and visual input are all collected into the NewsNet for the potential application in CV, NLP and Multi-modality.

2.2. Scope of Related Tasks

Since dense annotations and rich modalities are provided by NewsNet, we can conduct various applications on top of the datasets. Below, we show the parts of video tasks, which can be carried on the proposed dataset.

Common Video Tasks. The NewsNet is a high-quality video dataset, to evaluate the methods in common video tasks. For instance, NewsNet is an ideal source for video classification [10, 15, 37, 58, 59, 62] (including the out-of-distribution detection settings [49, 61, 72, 73] due to the existence of category 'Others' in *Topic* level). NewsNet holds annotations in four granularities for temporal segmentation. By training with the mentioned annotation, the method can **completely** split a video into several short segments, which will serve **partial** segmentation tasks like video grounding [40, 70, 74], captioning [3, 34, 71], retrieval [5, 16, 31, 52] and highlight detection [17, 18, 20, 28, 32, 54]. Meanwhile, we simultaneously provide aligned audio, text, and visual form in NewsNet, which might promote the improvement on the task of video grounding [30, 53, 66, 69], where the goal is to retrieve the corresponding video clip by adopting an agnostic or specific query as input. We also note that, over the last few years, the 'pre-training' and 'fine-tuning' paradigms [21, 24, 39, 46, 50, 65, 67, 68, 77] have become revolutionary in both NLP and CV communities. We believe that such technology will be further improved by introducing such a high-quality multi-modality video dataset.

Temporal Segmentation Approaches. Towards temporal video segmentation, existing works explore shot, scene and story segmentation tasks, respectively. Shot temporal segmentation [6, 11, 38, 55] is employed as a pre-processing to

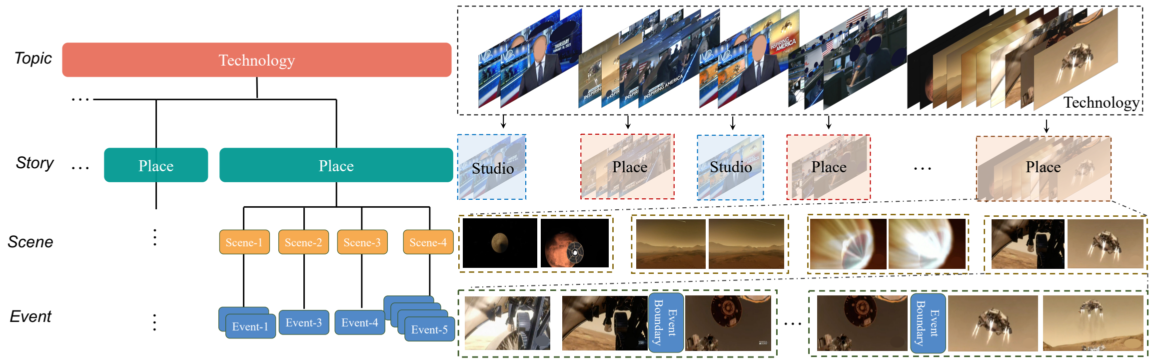

Figure 2. **Illustration of the four hierarchies.** The illustration shows a hierarchical diagram (*left*) along with an annotation example (*right*) for the four hierarchies, including *Topic*, *Story*, *Scene*, and *Event*. In the given sample, a technology (*Topic*) news video can be divided into several *Stories*, such as "studio" and "place". Taking "place" as an example, it can be further characterized into different *Scenes*, where the backgrounds or places are similar. Based on the changes of shots or persons in a scene, the *Scene* can be divided into atomic units, *i.e., Events*.

locate the transition positions in videos based on the similarity of the frames. For the scene segmentation, LGSS [47] converts the problem as a binary classification task at the shot level and applies a **boundary-based** model that aggregates adjacent shot features within a predefined sliding window to make a prediction. While SCRL [65] introduces a **boundary-free** model in a seq-to-seq way and reduces the inductive bias introduced by the predefined sliding window. As for story temporal segmentation, TI-NEWS [35] proposes a multi-modal framework based on the boundary-free model to complete the story segmentation and classification tasks simultaneously.

3. Dataset Summary

In this section, we will describe some details of the NewsNet. First, we will introduce the definition and taxonomy of the four hierarchical levels of semantics and show some annotated examples of them. Then, the statistical information of the proposed dataset will be given to better understand the data distribution of the NewsNet.

3.1. Definition and Taxonomy

Existing works focus on the shot- or event-level temporal segmentation, while neglecting the long-horizon cases, such as story- and topic-level segmentation. Hence this paper innovatively contributes more uniform benchmarks to the community, which covers all categories of temporal annotation. In order to clarify the meaning of the different labels, we detail them one by one:

- *Event*: the atom unit in our setting, which is defined by the switching of the shots or the changing of the person.
- *Scene*: a combination of several successive shots that focus on the same place from different angles. It is defined as a taxonomy-free unit.
- *Story*: a sequence of scenes about a piece of news broadcast in a studio or outdoors. Story-level segments can be categorized into 'Connection', 'Place' , 'Studio', 'Animation', 'Interview', and 'Photo'. A topic is composed of several stories.
- *Topic*: it summarizes the content of a long sequence, with the following keywords: 'Health', 'Politics', 'Entertainment', 'Economy', 'Crime', 'Weather', 'Sport', 'Technology', 'Military' and 'Others'. Generally, an individual news video contains several topics.

The above four different temporal annotations are defined in a hierarchical way. For example, a news video can be divided into several *Topic*-level segments; a *Topic* is further composed of several *Stories*; and so on. The difference between *Story* and *Scene* is that *Story* is category-aware while *Scene* is not, and the categories in *Story* are related to the data source. For example, *Story* contains a category of outdoor place footage, which is unique to news data and usually consists of multiple scenes, as illustrated in Fig. 2.

3.2. Collection and Statistics

As a large-scale dataset, we collect about 2,000 news broadcast videos from more than 20 different TV stations, of which 1,000 were annotated, and all the data are downloaded from the YouTube and Internet. As shown in Fig. 3, we show the distribution of program duration and sources. News is a kind of typical long-form video, ranging from 10 to 110 minutes. Most of the collected videos have a 1-hour duration, performing as a challenging span for temporal segmentation. During the annotation, we found that

Place refers to the news reported outside a studio.

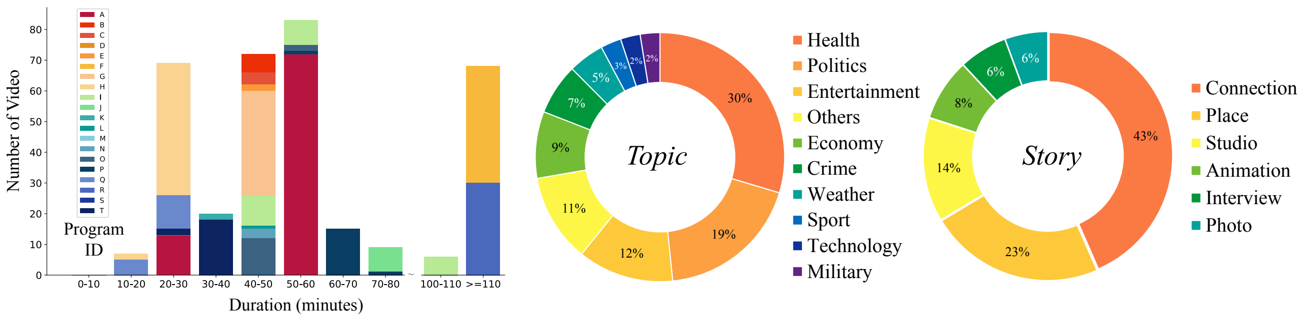

Figure 3. **The statistical information of NewsNet.** The left histogram shows the distribution of the Top-20 news programs, (*i.e.*, program A to T, by the total number of videos) in different duration, it can be found that the duration of the same program is relatively close. Besides, the two ring charts on the right show the category proportions of *Topic* and *Story*.

there exists a long-tailed distribution for the categories of the story- and topic-level annotation, leading to a practical challenge. We believe such a benchmark can simulate the real-world scenario effectively, so that the competing methods could be benchmarked in a practical manner. More details such as annotation process, visualization, *etc*., can be found in the *Supplementary Material*.

4. Benchmark and Protocol Navigation

In this section, we will first introduce two temporal segmentation task settings, *i.e.*, Separate Modeling and Hierarchical Modeling in NewsNet, with their definitions and protocols. Then, based on the potential relationship between different hierarchies, a simple yet effective loss function is introduced to better tackle hierarchical temporal segmentation. At last, we will briefly introduce the other common video tasks that can be conducted with NewsNet.

4.1. Hierarchical Temporal Segmentation

As long-form video can be divided into sub-video chunks according to its intrinsic properties, many researchers try to tackle the video segmentation problem from different perspectives in the last few years [12, 35, 47, 65]. Without loss of generality, we classify hierarchical temporal segmentation into two categories: *Separate Modeling* and *Hierarchical Modeling*. Separate modeling trains a segmentation model for each level segmentation task, while hierarchical modeling trains a model which simultaneously segments a video into multi-level segments.

NewsNet supports temporal segmentation tasks with various semantic granularity, *i.e.*, *Event*, *Scene*, *Story*, and *Topic*. Since many methods [12,47,65], take the *Event* (*aka Shot* sometimes) as the basic input unit to reduce the redundancy of the video and computational cost, NewsNet draws the video segmentation tasks on *Scene*, *Story*, and *Topic*. Moreover, these three hierarchies can be strictly aligned according to the *Event (Shot)* dimension, which is more conducive to *Hierarchical Modeling and more convenient for comparison*. Note that we also carried out the event-level experiment (see *Supplementary Material*).

Data and Modality. *In-domain*: all the videos are randomly split into training, validation, and testing sets with a ratio of 3:1:1. *Cross-domain*: training, validation, and testing sets are split according to different news program types/IDs; 5 different split combinations are set to ensure the generalization ability of the algorithm can be evaluated properly and fairly. As for the embeddings used in the experiment, the shot-level visual, audio, and textual features are extracted through ResNet50 [22] (pre-trained on Places365 [76]), public Bert-base model [14], and Cnn14 [27] (pre-trained on AudioSet [19]), respectively.

4.1.1 Separate Modeling

Task Definition. Given a video with the shot labels, the model requires to classify whether a shot is a segmentation point for one of the three semantic levels (*Scene*, *Story* and *Topic*). A segmentation point refers to a shot that lies at the end of a topic/story/scene segment, and its next shot is the start of a new topic/story/scene.

Backbone. In general, the common paradigm in those segmentation tasks is considering temporal segment to shot-level classification in boundary-base and -free way. We employ the state-of-the-art boundary-free model, SCRL [65], and the boundary-based model, LGSS [47] as our baseline approaches in separate modeling. More details about the protocol can be found in the *Supplementary Material*.

4.1.2 Hierarchical Modeling

Task Definition. Given a video with the shot labels, the model requires to classify whether a shot is a segmentation point for three semantic levels (*Scene*, *Story*, and *Topic*).

Backbone. In this setting, as shown in Fig. 4, we focus on two hierarchical modeling paradigms, including *Multi-Label* and *Multi-Head* with a novel learning objective called *Hierarchical Ranking*. More specifically, *Multi-*

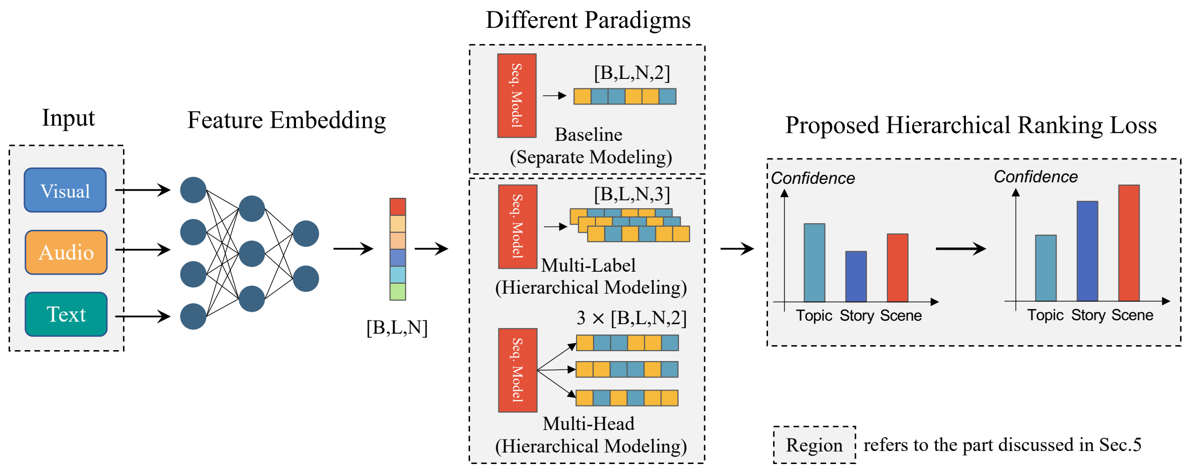

Figure 4. **An illustration of the separate modeling, two hierarchical modeling paradigms, and the proposed loss function.** For the baseline, the model is trained individually corresponding to different segmentation tasks. For hierarchical modeling, all the tasks are aggregated into one model. Specifically, B, L, and N represent the batch size, sequence length, and feature dimension, respectively.

Label works towards the hierarchical labels from the perspective of multi-label optimization, focusing on the aggregation of information into one head for estimation. While, *Multi-Head* addresses the challenge as a multi-task problem, where the tasks in different levels should be predicted via the individual pathways. Both pipelines are widely employed in modern architectures [13, 36, 60, 64], performing as simple but representative paradigms. Therefore, we adopt two hierarchical modeling paradigms as the baselines.

Hierarchical Ranking Loss. To better model the hierarchical temporal video segmentation task and combine the existing methods, we propose a new loss function called *Hierarchical Ranking Loss*. The motivation is: due to the accurate alignment between the hierarchical levels with the atom unit *Event*, as shown in Fig. 1, a positive segment boundary of the higher level task must be a positive segment boundary of the lower task. Hence, the segmentation confidence scores for different tasks should be ranked in a non-increasing order from coarse to fine for the paired task.

Given two tasks in the hierarchical modeling setting, *e.g.* task $\mathcal{H}$ and task $\mathcal{L}$, where the semantic level of task $\mathcal{H}$ is higher than that of task $\mathcal{L}$, the proposed hierarchical ranking loss function is as follows:

$$\mathcal{P}_{\mathcal{H},\mathcal{L}} = \frac{1}{B}\frac{1}{L}\cdot\sum[\sigma(F_{\mathcal{H}} - \mathcal{SG}(F_{\mathcal{L}}))\cdot Y_{\mathcal{L}}] \quad (1)$$

where B and L are the batch size and sequence length respectively; $\sigma(\cdot)$ refers to the contraction mapping, like sigmoid function; F stands for the segmentation confidence score of the task; $\mathcal{SG}$ is the *Stop Gradient* operation for the computational graph, and $Y_{\mathcal{L}}$ corresponds to the label of the task $\mathcal{L}$.

And the final objective function for all three tasks can be formed as follow:

$$\mathcal{L} = \mathcal{P}_{Topic,Story} + \mathcal{P}_{Topic,Scene} + \mathcal{P}_{Story,Scene} + \mathcal{L}_{CE} \quad (2)$$

where $\mathcal{L}_{CE}$ is the cross-entropy loss function.

4.2. Common Video Understanding Tasks

The NewsNet could also apply other common video understanding tasks, *e.g.*, video classification, localization, and highlight detection. More details can be found in *Supplementary Material*.

5. Experiment and Analysis

In this section, we mainly focus on the temporal segmentation tasks on the NewsNet and further contribute several benchmarks on the different baselines. In particular, we conduct the empirical study to thoroughly benchmark the effectiveness of model structure, multi-modality information, domain gap, and hierarchical modeling. Note that, due to the limitation on page length, the implementation details and the results for other common video tasks are given in the *Supplementary Material*.

Table 2. The performance between boundary-based (B.B.) and boundary-free (B.F) model with visual input on F1 score ↑, Precision ↑, and Recall↑. Light gray refers to the failure case.

Task	Model	F1 score	Precision	Recall
Scene	B.B.	77.3	67.7	90.2
	B.F.	76.8	76.1	77.5
Story	B.B.	70.7	65.9	76.2
	B.F.	71.2	72.3	70.0
Topic	B.B.	13.1	7.0	100.0
	B.F.	62.9	72.4	55.6

As an overview, based on the large-scale dataset with rich modalities, we conduct several experiments to respond to five critical questions concerned by the community:

Table 3. In-domain performance by using boundary-free (B.F.) model. The **bolded** values stand for the optimal performances for each task.

Task	Modality	F1 score	Precision	Recall
Scene	V	76.8	76.1	77.5
	A	69.8	66.8	73.2
	T	66.7	56.3	**81.9**
	V+A+T	**78.3**	**80.9**	75.8
Story	V	71.2	72.3	70.0
	A	59.3	57.6	61.1
	T	50.6	57.4	45.2
	V+A+T	**75.4**	**74.7**	**76.2**
Topic	V	62.9	72.4	55.6
	A	58.1	59.4	56.9
	T	39.0	46.5	33.5
	V+A+T	**73.2**	**74.3**	**72.2**

Table 4. Cross-domain setting by using boundary-free (B.F.) model. The **bolded** values stand for the optimal performances for each task.

Task	Modality	Avg. F1 score (std.)	Avg. Precision	Avg. Recall
Scene	V	72.9 (2.1)	70.9	75.2
	A	62.7 (4.0)	59.7	66.6
	T	61.6 (5.0)	52.8	77.0
	V+A+T	**76.0** (2.1)	**74.4**	**77.9**
Story	V	68.5 (2.6)	70.3	66.9
	A	55.7 (3.6)	53.6	59.0
	T	51.1 (3.6)	43.4	65.4
	V+A+T	**72.9** (2.2)	**73.7**	**72.4**
Topic	V	60.6 (4.7)	69.8	53.8
	A	59.0 (5.2)	56.0	62.9
	T	49.8 (5.2)	45.7	55.9
	V+A+T	**72.2** (3.6)	**72.3**	**72.5**

1. Can existing methods also perform well on long-horizon temporal segmentation tasks, such as topic-level and story-level segmentation?
2. Do the different-level tasks test the ability of the model on viewing different temporal spans?
3. Can existing methods well address the semantic gap between low- and high-level granularities?
4. Will the temporal segmentation models be benefited from multi-modality information?
5. How could the models benefit from Hierarchical Modeling?

5.1. Results against Increased Temporal Span

To investigate the generalization of the existing methods against the long-horizon cases, we conduct an experiment on two different head architectures, including boundary-free and boundary-based heads, on three tasks: scene-level, story-level, and topic-level temporal segmentation. As shown in Fig. 2, the performance of the boundary-based model has a significant drop on the *Topic* task. Compared with the other tasks, the *Topic* segmentation task has more time slices for input, leading to a hard challenge caused by longer time dependencies. The methods can segment scenes with about 76-77 F1 score while degrading to 62.9 on Topic-level segmentation. For the first question, the answer is that *both the boundary-free model and the boundary-based model can not generalize well in the story- and topic-level segmentation*. Combining the overall performance and stability on different tasks, the next experiments will utilize a boundary-free model as the basic backbone under comprehensive consideration.

5.2. Performance against Cross News Domains

To investigate the generalization of the existing method, we analyze the performance of the method against the varieties of different news programs. As shown in Table 3 and 4, the cross-domain setting slightly degrades the performance on all the tasks. Comparing the in-domain results using the full modality in Table 3 and corresponding cross-domain results in Table 4, the F1 score decreases by 2.3 on task *Scene*, by 2.5 on task *Story*, and by 1.0 on task *Topic*. The reason for this result may be that higher-level tasks, such as *Topic*, has more abstract semantics and the knowledge is easier to transfer. In contrast, task *Scene* in the lowest level, generally only considers place switching as the segmentation cue. Besides, the category of *Scene* is an open set, leading to a greater challenge in the cross-domain setting. Hence, *temporal segmentation model also suffers from the out-of-distribution case.* Based on the property of the collected dataset, we set the videos from different news programs as different domains, and establish a comprehensive benchmark to facilitate the research on such line.

5.3. Performance using Multi-Modalities

When it comes to multi-modality, as shown in Table. 4, *the model can be benefited from multimodal information among all the tasks*. For example, the performance of using only visual modality achieves only 60.6 recall in topic-level segmentation. However, by aggregating the other modalities including audio and text, the performance can be improved significantly (72.2 vs 49.8), which indicates the potential value for addressing long-discrepancy via multi-modality technology. Note that, the improvement in *Scene* segmentation is, to some extent, limited, we believe that *Scene* is mainly defined by visual input, hence can not obtain adequate complementarity from other modalities.

5.4. Performance based on Hierarchical Modeling

Here, we discuss the core challenge revealed in this paper: 'How could the model benefit from hierarchical annotations?'. In this case, the baseline is set as the model, which can only be accessible to one specific-level annotation. As

Table 5. The F1 scores of baselines trained with different levels of annotations on full modalities without our hierarchical ranking loss, where blue and orange indicate the in-domain and cross-domain, respectively. Each row refers to the result corresponding to a single task. Hie. Modeling stands for Hierarchical Modeling while Sep. Modeling is Separate Modeling.

Recipe	Baseline Sep. Modeling	Multi-Label Hie. Modeling	Multi-Head Hie. Modeling
Scene	78.3 / 76.0	79.1 / 76.5	**79.9** / **76.9**
+ Story	75.4 / 72.9	**75.4** / **74.7**	74.2 / 74.0
Scene	78.3 / 76.0	**79.8** / 76.4	79.5 / **76.5**
+ Topic	**73.2** / 72.2	70.5 / 72.8	70.9 / **73.0**
Story	75.4 / 72.9	**76.2** / **74.3**	75.4 / 73.9
+ Topic	73.2 / 72.2	**77.3** / 73.2	75.2 / **73.5**
Scene	78.3 / 76.0	77.4 / **76.8**	**79.8** / **76.8**
+ Story	**75.4** / 72.9	74.3 / **74.3**	74.5 / 73.7
+ Topic	73.2 / 72.2	74.3 / **72.6**	**76.6** / 70.4

listed in Table 5, we consider two simple strategies, including *Multi-Label* and *Multi-Head*, to aggregate the hierarchical information. By training with multi-level annotations, a $1.0 \sim 3.0$ % improvement can be achieved, which indicates, the segmentation can be benefited by introducing the hierarchical structures of videos. However, Table 5 also shows directly combining different-level granularities might not result in gains due to the semantic gap between low- and high-level granularities. For example, directly combining scene- and topic-level information by Multi-Label hierarchical modeling or Multi-Head hierarchical modeling does not consistently improve the topic-level segmentation performance. This poses a new and challenging research question, *i.e*. how to jointly train with various granularities toward hierarchical modeling of long-form videos?

To address the above question, we introduce a simple but effective loss function, *i.e., Hierarchical Ranking Loss* introduced in Sec. 4.1.2. As shown in Table 6, by combining with our proposed loss, the F1 score for segmentation can be further improved in most cases. Typically, in *Topic*-level, Multi-Head coupled with the proposed loss can reach the best performance, *i.e.*, improving Multi-Head case from 70.4 to 73.2 F1 score.

5.5. Other Benchmarks involved in the NewsNet

In addition to the temporal segmentation tasks, we also conduct extensive common video tasks, including highlight detection, video classification, and video localization, on the proposed dataset. Due to the limited pages, we show this part in the *Supplementary Materials*.

Table 6. The F1 scores of the methods with or without hierarchical ranking loss under the in-domain / cross-domain setting on full modalities. Hie. stands for Hierarchical Modeling and Sep. refers to Separate Modeling.

Method	Scene	Story	Topic
Baseline (Sep.)	78.3 / 76.0	75.4 / 72.9	73.2 / 72.2
Multi-Label (Hie.)	77.4 / 76.8	74.3 / **74.3**	74.3 / 72.6
Multi-Label w/ Hie. Loss (Hie.)	**79.6** / **76.9**	**74.4** / 73.5	**77.8** / **73.1**
Multi-Head (Hie.)	79.8 / 76.8	74.5 / 73.7	**76.6** / 70.4
Multi-Head w/ Hie. Loss (Hie.)	**80.3** / **76.9**	**76.3** / **74.6**	76.5 / **73.2**

6. Conclusion, Limitation and Future work

We introduce NewsNet, a new large-scale dataset for temporal video segmentation. Compared with existing large-scale datasets for temporal video segmentation, NewsNet additionally provides two abstractive levels of temporal segmentation, which have not been taken into account by these datasets. The four-level and hierarchical annotations enable the community to explore how to comprehensively represent the complex structure of long-form videos from coarse to fine. Extensive efforts of human workers are devoted, so as to ensure that annotations are high-quality. We hope that Newsnet and baseline results can facilitate the development of temporal video segmentation in terms of insightful research and practical tools. In addition, as NewsNet provides diverse annotations for dividing a video into segments of different granularities as well as category labels, it can also serve other video understanding tasks such as video grounding, captioning, retrieval, highlight detection, and video classification.

In this paper, we only focus on the problem of hierarchical temporal segmentation. We have not explored other multi-modality tasks, such as visual reasoning and video question answering, while our dataset can serve these tasks thanks to its multi-modality nature. We will explore these tasks in our future work to further promote advances in multi-modality-based video understanding and reasoning.

Acknowledgement

This work was supported by the KAUST Office of Sponsored Research through the Visual Computing Center (VCC) funding, as well as, the SDAIA-KAUST Center of Excellence in Data Science and Artificial Intelligence (SDAIA-KAUST AI). Raghavendra Ramachandra is supported under the SALT and OFFPAD projects funded by the Research Council of Norway, Norway. Chia-Wen Lin is supported by National Science and Technology Council, Taiwan, under project NSTC 111-2634-F-002-023.

References

[1] LOVEU21. https://sites.google.com/view/loveucvpr21, 2021. 3

[2] LOVEU22 program schedule. https://sites.google.com/view/loveucvpr22/program, 2022. 3

[3] Nayyer Aafaq, Naveed Akhtar, Wei Liu, Syed Zulqarnain Gilani, and Ajmal Mian. Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12487–12496, 2019. 2, 3

[4] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016. 2, 3

[5] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021. 2, 3

[6] Lorenzo Baraldi, Costantino Grana, and Rita Cucchiara. A deep siamese network for scene detection in broadcast videos. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1199–1202, 2015. 2, 3

[7] Lorenzo Baraldi, Costantino Grana, and Rita Cucchiara. Shot and scene detection via hierarchical clustering for re-using broadcast video. In *International conference on computer analysis of images and patterns*, pages 801–811. Springer, 2015. 2, 3

[8] Joe Barrow, Rajiv Jain, Vlad Morariu, Varun Manjunatha, Douglas W Oard, and Philip Resnik. A joint model for document segmentation and segment labeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 313–322, 2020. 2

[9] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015. 3

[10] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 1, 3

[11] Vasileios T Chasanis, Aristidis C Likas, and Nikolaos P Galatsanos. Scene detection in videos using shot clustering and sequence alignment. *IEEE transactions on multimedia*, 11(1):89–100, 2008. 3

[12] Shixing Chen, Xiaohan Nie, David Fan, Dongqing Zhang, Vimal Bhat, and Raffay Hamid. Shot contrastive self-supervised learning for scene boundary detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9796–9805, 2021. 2, 5

[13] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 6

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 5

[15] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 1, 3

[16] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *European Conference on Computer Vision*, pages 214–229. Springer, 2020. 2, 3

[17] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275, 2017. 3

[18] Ana Garcia del Molino and Michael Gygli. Phd-gifs: personalized highlight detection for automatic gif creation. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 600–608, 2018. 3

[19] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017. 5

[20] Michael Gygli, Yale Song, and Liangliang Cao. Video2gif: Automatic generation of animated gifs from video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1001–1009, 2016. 3

[21] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 3

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[23] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie understanding. In *European Conference on Computer Vision*, pages 709–727. Springer, 2020. 2, 3

[24] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. *arXiv preprint arXiv:2112.04478*, 2021. 3

[25] Hyolim Kang, Jinwoo Kim, Taehyun Kim, and Seon Joo Kim. Uboco: Unsupervised boundary contrastive learning for generic event boundary detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20073–20082, 2022. 2, 3

[26] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 2, 3

[27] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020. 5

[28] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715, 2017. 3

[29] Kathy Lee, Diana Palsetia, Ramanathan Narayanan, Md Mostofa Ali Patwary, Ankit Agrawal, and Alok Choudhary. Twitter trending topic classification. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 251–258. IEEE, 2011. 2

[30] Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34:11846–11858, 2021. 3

[31] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7331–7341, 2021. 2, 3

[32] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *European Conference on Computer Vision*, pages 447–463. Springer, 2020. 3

[33] Congcong Li, Xinyao Wang, Longyin Wen, Dexiang Hong, Tiejian Luo, and Libo Zhang. End-to-end compressed video representation learning for generic event boundary detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13967–13976, 2022. 3

[34] Kevin Lin, Linjie Li, Chung-Ching Lin, Faisal Ahmed, Zhe Gan, Zicheng Liu, Yumao Lu, and Lijuan Wang. Swinbert: End-to-end transformers with sparse attention for video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17949–17958, 2022. 2, 3

[35] Ye Liu, Lingfeng Qiao, Di Yin, Zhuoxuan Jiang, Xinghua Jiang, Deqiang Jiang, and Bo Ren. Os-msl: One stage multimodal sequential link framework for scene segmentation and classification. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6269–6277, 2022. 2, 3, 4, 5

[36] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, October 2021. 6

[37] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3202–3211, 2022. 1, 3

[38] Jakub Lokoč, Gregor Kovalčík, Tomáš Souček, Jaroslav Moravec, and Přemysl Čech. A framework for effective known-item search in video. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1777–1785, 2019. 2, 3

[39] Sauradip Nag, Xiatian Zhu, Yi-Zhe Song, and Tao Xiang. Zero-shot temporal action detection via vision-language prompting. *arXiv preprint arXiv:2207.08184*, 2022. 3

[40] Guoshun Nan, Rui Qiao, Yao Xiao, Jun Liu, Sicong Leng, Hao Zhang, and Wei Lu. Interventional video grounding with dual contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2765–2775, 2021. 3

[41] Fahmi Salman Nurfikri, Mohamad Syahrul Mubarok, et al. News topic classification using mutual information and bayesian network. In *2018 6th International Conference on Information and Communication Technology (ICoICT)*, pages 162–166. IEEE, 2018. 2

[42] Sofia Ares Oliveira, Benoit Seguin, and Frederic Kaplan. dhsegment: A generic deep-learning approach for document segmentation. In *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 7–12. IEEE, 2018. 2

[43] Alejandro Pardo, Fabian Caba Heilbron, Juan León Alcázar, Ali Thabet, and Bernard Ghanem. Moviecuts: A new dataset and benchmark for cut type recognition. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, pages 668–685, 2022. 3

[44] Will Price, Carl Vondrick, and Dima Damen. Unweavenet: Unweaving activity stories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13770–13779, 2022. 3

[45] Zhiwu Qing, Shiwei Zhang, Ziyuan Huang, Yi Xu, Xiang Wang, Mingqian Tang, Changxin Gao, Rong Jin, and Nong Sang. Learning from untrimmed videos: Self-supervised video representation learning with hierarchical consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13821–13831, 2022. 3

[46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 3

[47] Anyi Rao, Linning Xu, Yu Xiong, Guodong Xu, Qingqiu Huang, Bolei Zhou, and Dahua Lin. A local-to-global approach to multi-modal movie scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10146–10155, 2020. 2, 3, 4, 5

[48] Daniel Rotman, Dror Porat, and Gal Ashour. Optimal sequential grouping for robust video scene detection using multiple modalities. *International Journal of Semantic Computing*, 11(02):193–208, 2017. 2, 3

[49] Ellen Rushe and Brian Mac Namee. Anomaly detection in raw audio using deep autoregressive networks. In *ICASSP 2019-2019 IEEE International Conference on Acoustics,*

Speech and Signal Processing (ICASSP), pages 3597–3601. IEEE, 2019. 3

[50] Weimin Shi, Mingchen Zhuge, Zhong Zhou, Dehong Gao, and Deng-Ping Fan. Qr-clip: Introducing explicit open-world knowledge for location and time reasoning. *arXiv preprint arXiv:2302.00952*, 2023. 3

[51] 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*, pages 8075–8084, 2021. 2, 3

[52] Nina Shvetsova, Brian Chen, Andrew Rouditchenko, Samuel Thomas, Brian Kingsbury, Rogerio S Feris, David Harwath, James Glass, and Hilde Kuehne. Everything at once-multi-modal fusion transformer for video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20020–20029, 2022. 2, 3

[53] Mattia Soldan, Alejandro Pardo, Juan León Alcázar, Fabian Caba, Chen Zhao, Silvio Giancola, and Bernard Ghanem. Mad: A scalable dataset for language grounding in videos from movie audio descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5026–5035, 2022. 3

[54] Yale Song, Miriam Redi, Jordi Vallmitjana, and Alejandro Jaimes. To click or not to click: Automatic selection of beautiful thumbnails from videos. In *Proceedings of the 25th ACM international on conference on information and knowledge management*, pages 659–668, 2016. 3

[55] Tomáš Souček and Jakub Lokoč. Transnet v2: an effective deep network architecture for fast shot transition detection. *arXiv preprint arXiv:2008.04838*, 2020. 2, 3

[56] Jiaqi Tang, Zhaoyang Liu, Chen Qian, Wayne Wu, and Limin Wang. Progressive attention on multi-level dense difference maps for generic event boundary detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3355–3364, 2022. 2, 3

[57] Shitao Tang, Litong Feng, Zhanghui Kuang, Yimin Chen, and Wei Zhang. Fast video shot transition localization with deep structured models. In *Asian Conference on Computer Vision*, pages 577–592. Springer, 2018. 2

[58] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 1, 3

[59] Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. Video classification with channel-separated convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5552–5561, 2019. 1, 3

[60] Dongkai Wang and Shiliang Zhang. Unsupervised person re-identification via multi-label classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 6

[61] Jue Wang and Anoop Cherian. Gods: Generalized one-class discriminative subspaces for anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8201–8211, 2019. 3

[62] 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 *European conference on computer vision*, pages 20–36. Springer, 2016. 1, 3

[63] Yuxuan Wang, Difei Gao, Licheng Yu, Weixian Lei, Matt Feiszli, and Mike Zheng Shou. Geb+: A benchmark for generic event boundary captioning, grounding and retrieval. In *European Conference on Computer Vision*, pages 709–725. Springer, 2022. 2

[64] Jonatas Wehrmann, Ricardo Cerri, and Rodrigo Barros. Hierarchical multi-label classification networks. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5075–5084. PMLR, 10–15 Jul 2018. 6

[65] Haoqian Wu, Keyu Chen, Yanan Luo, Ruizhi Qiao, Bo Ren, Haozhe Liu, Weicheng Xie, and Linlin Shen. Scene consistency representation learning for video scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14021–14030, 2022. 2, 3, 4, 5

[66] Bo Xiong, Yannis Kalantidis, Deepti Ghadiyaram, and Kristen Grauman. Less is more: Learning highlight detection from video duration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1258–1267, 2019. 3

[67] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021. 3

[68] Mengmeng Xu, Juan-Manuel Pérez-Rúa, Victor Escorcia, Brais Martinez, Xiatian Zhu, Li Zhang, Bernard Ghanem, and Tao Xiang. Boundary-sensitive pre-training for temporal localization in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7220–7230, 2021. 3

[69] Minghao Xu, Hang Wang, Bingbing Ni, Riheng Zhu, Zhenbang Sun, and Changhu Wang. Cross-category video highlight detection via set-based learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7970–7979, 2021. 3

[70] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Tubedetr: Spatio-temporal video grounding with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16442–16453, 2022. 3

[71] Hanhua Ye, Guorong Li, Yuankai Qi, Shuhui Wang, Qingming Huang, and Ming-Hsuan Yang. Hierarchical modular network for video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17939–17948, 2022. 2, 3

[72] Muhammad Zaigham Zaheer, Arif Mahmood, Marcella Astrid, and Seung-Ik Lee. Claws: Clustering assisted weakly supervised learning with normalcy suppression for anomalous event detection. In *European Conference on Computer Vision*, pages 358–376. Springer, 2020. 3

[73] M Zaigham Zaheer, Arif Mahmood, M Haris Khan, Mattia Segu, Fisher Yu, and Seung-Ik Lee. Generative cooperative learning for unsupervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14744–14754, 2022. 3

[74] Runhao Zeng, Haoming Xu, Wenbing Huang, Peihao Chen, Mingkui Tan, and Chuang Gan. Dense regression network for video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10287–10296, 2020. 3

[75] Haoxin Zhang, Zhimin Li, and Qinglin Lu. Better learning shot boundary detection via multi-task. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4730–4734, 2021. 2, 3

[76] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017. 5

[77] Mingchen Zhuge, Dehong Gao, Deng-Ping Fan, Linbo Jin, Ben Chen, Haoming Zhou, Minghui Qiu, and Ling Shao. Kaleido-bert: Vision-language pre-training on fashion domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12647–12657, 2021. 3