NewsNet: A Novel Dataset for Hierarchical Temporal Segmentation

Haoqian Wu†, Keyu Chen†, Haozhe Liu†, Mingchen Zhuge†, Bing Li‡, Ruizhi Qiao†, Xiujun Shu†, Bei Gan†, Liangsheng Xu†, Bo Ren†, Mengmeng Xu‡, Wentian Zhang‡, Raghavendra Ramachandra§, Chia-Wen Lin¶, Bernard Ghanem‡

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)

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

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].
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.

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.)

<table>
<thead>
<tr>
<th>Dataset</th>
<th># Video</th>
<th>Duration (hours)</th>
<th>Modality</th>
<th># Annotation(s) per Video</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Topic</td>
<td>Story</td>
</tr>
<tr>
<td>AVS [75]</td>
<td>197</td>
<td>-</td>
<td>Visual</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>OVSD [48]</td>
<td>21</td>
<td>10</td>
<td>Visual</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Kinetics-GEBD [51]</td>
<td>54,691</td>
<td>152</td>
<td>Audio + Visual</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MovieNet [23]</td>
<td>1,100</td>
<td>2174</td>
<td>Text + Audio + Visual</td>
<td>-</td>
<td>66.0</td>
</tr>
<tr>
<td>RAI [7]</td>
<td>10</td>
<td>-</td>
<td>Visual</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>TI-News [35]</td>
<td>477</td>
<td>244</td>
<td>Audio + Visual</td>
<td>-</td>
<td>55.6</td>
</tr>
<tr>
<td>NewsNet (Ours)</td>
<td>1,000</td>
<td>946</td>
<td>Text + Audio + Visual</td>
<td>8.5</td>
<td>51.6</td>
</tr>
</tbody>
</table>

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


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.*
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.


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-
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 $H$ and task $L$, where the semantic level of task $H$ is higher than that of task $L$, the proposed hierarchical ranking loss function is as follows:

$$\mathcal{P}_{H,L} = \frac{1}{B \cdot L} \sum \left[ \sigma(F_H - S\nabla(F_L)) \cdot Y_L \right]$$  \hspace{1cm} (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; $S\nabla$ is the Stop Gradient operation for the computational graph, and $Y_L$ corresponds to the label of the task $L$.

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

$$\mathcal{L} = \mathcal{P}_{\text{Topic,Story}} + \mathcal{P}_{\text{Topic,Scene}} + \mathcal{P}_{\text{Story,Scene}} + \mathcal{L}_{CE}$$  \hspace{1cm} (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.

<table>
<thead>
<tr>
<th>Task</th>
<th>Model</th>
<th>F1 score</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scene</td>
<td>B.B.</td>
<td>77.3</td>
<td>67.7</td>
<td>90.2</td>
</tr>
<tr>
<td></td>
<td>B.F.</td>
<td>76.8</td>
<td>76.1</td>
<td>77.5</td>
</tr>
<tr>
<td>Story</td>
<td>B.B.</td>
<td>70.7</td>
<td>65.9</td>
<td>76.2</td>
</tr>
<tr>
<td></td>
<td>B.F.</td>
<td>71.2</td>
<td>72.3</td>
<td>70.0</td>
</tr>
<tr>
<td>Topic</td>
<td>B.B.</td>
<td>13.1</td>
<td>7.0</td>
<td>100.0</td>
</tr>
<tr>
<td></td>
<td>B.F.</td>
<td>62.9</td>
<td>72.4</td>
<td>55.6</td>
</tr>
</tbody>
</table>

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.

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

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 multimodality 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
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.

<table>
<thead>
<tr>
<th>Recipe</th>
<th>Baseline (Sep.)</th>
<th>Multi-Label (Hie.)</th>
<th>Multi-Head (Hie.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scene</td>
<td>78.3 / 76.0</td>
<td>79.1 / 76.5</td>
<td>79.9 / 76.9</td>
</tr>
<tr>
<td>+ Story</td>
<td>75.4 / 72.9</td>
<td>75.4 / 74.7</td>
<td>74.0 / 74.0</td>
</tr>
<tr>
<td>Scene</td>
<td>78.3 / 76.0</td>
<td>79.8 / 76.4</td>
<td>79.5 / 76.5</td>
</tr>
<tr>
<td>+ Topic</td>
<td>73.2 / 72.2</td>
<td>70.5 / 72.8</td>
<td>70.9 / 73.0</td>
</tr>
<tr>
<td>Story</td>
<td>75.4 / 72.9</td>
<td>76.2 / 74.3</td>
<td>75.4 / 73.9</td>
</tr>
<tr>
<td>+ Topic</td>
<td>73.2 / 72.2</td>
<td>77.3 / 73.2</td>
<td>75.2 / 73.5</td>
</tr>
<tr>
<td>Scene</td>
<td>78.3 / 76.0</td>
<td>77.4 / 76.8</td>
<td>79.8 / 76.8</td>
</tr>
<tr>
<td>+ Story</td>
<td>75.4 / 72.9</td>
<td>74.3 / 74.3</td>
<td>74.5 / 73.7</td>
</tr>
<tr>
<td>+ Topic</td>
<td>73.2 / 72.2</td>
<td>74.3 / 72.6</td>
<td>76.6 / 70.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Scene</th>
<th>Story</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (Sep.)</td>
<td>78.3 / 76.0</td>
<td>75.4 / 72.9</td>
<td>71.2 / 72.2</td>
</tr>
<tr>
<td>Multi-Label (Hie.)</td>
<td>77.4 / 76.8</td>
<td>74.3 / 74.3</td>
<td>74.3 / 72.6</td>
</tr>
<tr>
<td>Multi-Label w/ Hie. Loss (Hie.)</td>
<td>79.6 / 76.9</td>
<td>74.4 / 73.5</td>
<td>77.8 / 73.1</td>
</tr>
<tr>
<td>Multi-Head (Hie.)</td>
<td>79.8 / 76.8</td>
<td>74.5 / 73.7</td>
<td>76.6 / 70.4</td>
</tr>
<tr>
<td>Multi-Head w/ Hie. Loss (Hie.)</td>
<td>80.3 / 76.9</td>
<td>76.3 / 74.6</td>
<td>76.5 / 73.2</td>
</tr>
</tbody>
</table>

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. Chi-Wen Lin is supported by National Science and Technology Council, Taiwan, under project NSTC 111-2634-F-002-023.

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.
References

[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


