Supplementary Material
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)

Abstract

In this supplementary, we provide additional detailed information about our NewsNet in Sec. A, where we elaborate on annotation details, and describe the proposed annotation pipeline and the annotation User Interface (UI). Moreover, we present more details about our hierarchical ranking loss in Sec. B.

We also provide more experimental results in Sec. C. Besides temporal video segmentation, our dataset can be applied to various tasks such as video classification, video highlight detection, and video localization, thanks to the rich annotations of NewsNet. We construct baselines on top of state-of-the-art approaches in these tasks, showing our dataset can serve several video understanding areas.

A. More Details of NewsNet Annotations

In this section, we present more details about the annotations of our NewsNet and then visualize some annotation results for illustration.

A.1. Temporal Segmentation Annotations

NewsNet aims to provide four hierarchical levels of annotations for temporal video segmentation, according to different levels of semantics in terms of topic, story, scene and event. However, it is difficult and excessively time-consuming for non-expert workers to directly annotate such four levels of annotations, due to the complex content of long-form videos and the ambiguity of segment boundaries. For example, it is crucial for human workers to understand the difference between different levels of semantics, such that they can annotate the temporal structure in a four-level hierarchy. Furthermore, some workers may introduce inconsistent annotations without clear and thorough instructions.

In this paper, instead of spending much time training workers, we propose a new annotation scheme by exploiting domain knowledge of news videos, which helps human workers annotate temporal segmentation efficiently. Moreover, we develop an annotation platform with User Interface (UI) by which workers can conveniently annotate long-form videos and review others' annotation results.

Annotation Scheme. We develop a novel annotation scheme for NewsNet by exploiting the properties of news videos, to help workers annotate four-level temporal structures of long-form videos efficiently.

For *Event*-level annotation, we utilize the shot detection algorithm [9] to find boundaries of shots and employ a person search algorithm [10] to detect subject changes. However, such annotation results are often not completely accurate, especially on challenging videos, due to the limitations of the algorithms. Workers further manually refine these event-level annotation results. Based on refined event-level annotations, the scene-, story- and topic-level annotations are generated using the proposed annotation scheme illustrated in Fig. B. The scheme exploits the properties of news videos to help workers with annotation. For example, since a story segment falls under a topic segment while scene falls under story, a worker can recursively annotate story and scene segments after detecting the beginning of a topic segment. In addition, the transition (*i.e.,* a segment of video animation used to bridge the content between two shots) and advertisement (ad) segments are treated as outliers in our paper, since the content of transitions and ads are unrelated to news. Nevertheless, transitions and ads provide discriminative information which indicates the boundaries

† Equal Contribution

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

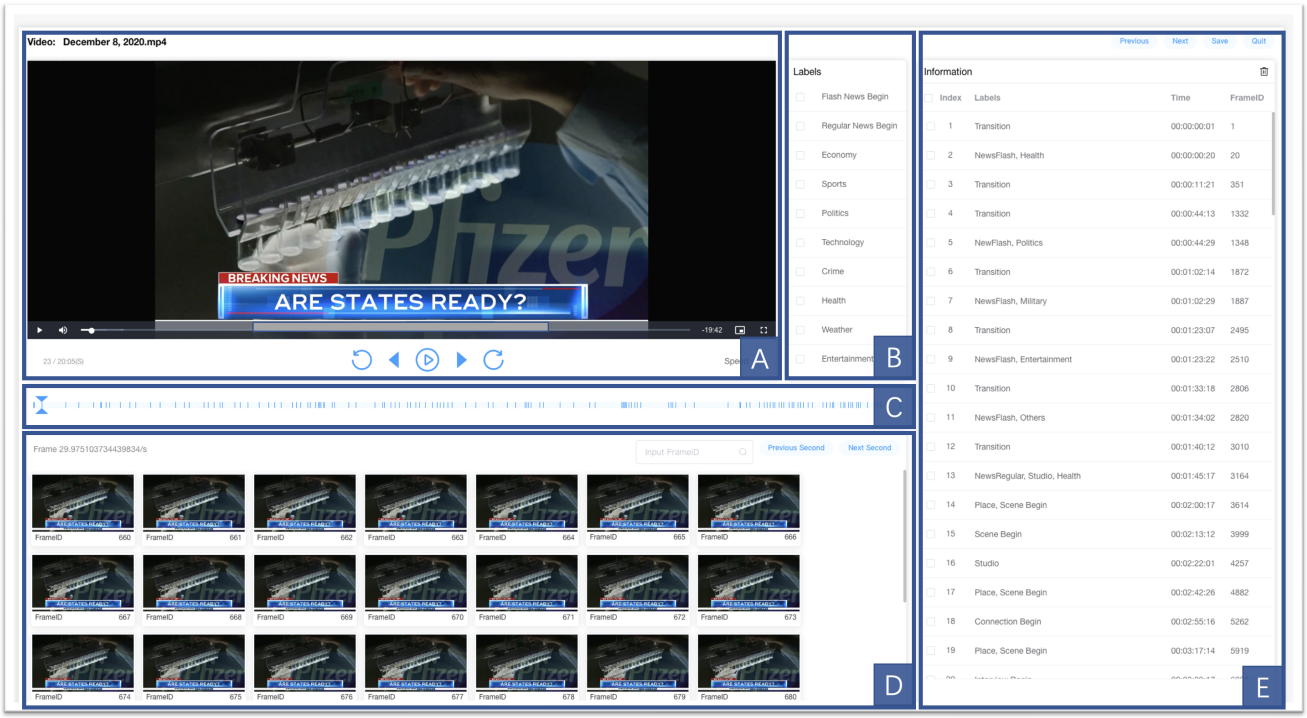

Figure A. **The annotation UI for annotators.** *Zone A* is used to display the video picture and control the playback progress of the video; All the labels are presented in *Zone B*; All labeled timestamps are presented on the timeline in *Zone C; Zone D* is used to present all the frames in a certain second, which is convenient for annotators to find the exact segmentation boundary; *Zone E* is used to display all labeled results, timestamps and frame IDs, and users can jump to the corresponding pictures in *Zone A* by clicking on a specific item.

of segments (*e.g.* topic-level segments).

Annotation UI. We design an annotation platform with a user-friendly UI for human workers to annotate videos and review annotation results. Fig. A shows the developed UI. The platform supports dense annotations at the frame level and allows us to annotate different levels of temporal structures hierarchically using the pipeline shown in Fig. B.

A.2. Additional Annotations

Video Highlight Annotations. Our NewsNet provides additional annotations for video highlight detection, thanks to the properties of news videos. In particular, given a piece of news, video producers often place news flash or breaking news at the beginning for a brief introduction, and then broadcast the detailed information of the news, where we refer to the detailed video content as regular news. In other words, **the news flash can be considered as the highlight of the corresponding regular news**. Hence, we pair each news flash and its corresponding regular news to provide video highlight annotations. As such, NewsNet can cover highlight annotation tasks.

A.3. Visualization of the Annotation

We show two annotation examples in Fig. C, the colored vertical bar is generated by the ground-truth annotations.

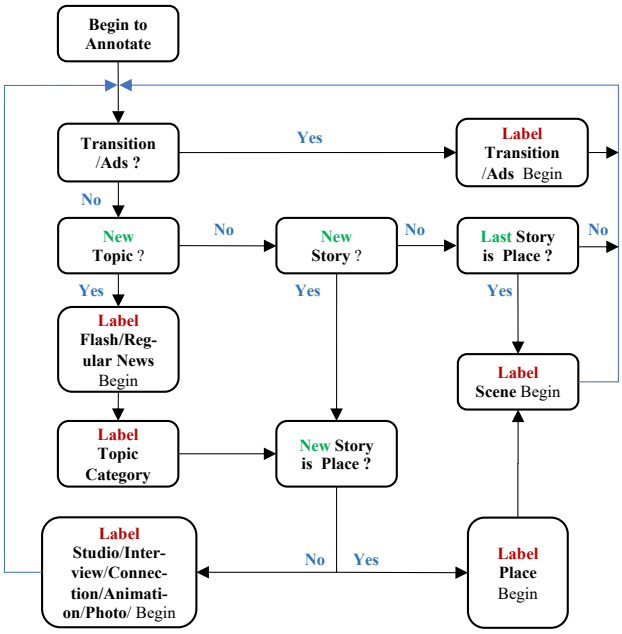

Figure B. **Illustration of our annotation scheme.**

A.4. Societal Impact and Privacy

In the collecting of the dataset, we have followed several rules to ensure the preservation of privacy:

- We mask all the faces which appear in the video, while mosaicing all the objects related to personal information.
- We only adopt videos published on the Internet for non-commercial use.
- We chuck out videos related to sensitive topics such as racism, militarism, and violence.

B. More Details of Hierarchical Ranking Loss

In the Sec. 4.1.2 of the main paper, we have discussed the proposed *Hierarchical Ranking Loss*. Here we introduce more information about the loss and provide more experimental analyses.

PyTorch-like Pseudo Code. In order to describe the implementation of the loss function, a piece of pseudo-code is utilized to describe its calculation process (see Listing 1). The proposed loss function can be plug-and-play with regular cross-entropy loss, achieving significant improvement at a negligible computational cost on our NewsNet.

```
import torch as T

def Hierarchical_Ranking_Loss(out, label):
    f_sc = out["scene"]
    f_st = out["story"]
    f_to = out["topic"]
    lab_sc = label["scene"]
    lab_st = label["story"]

    out_st_sc = T.sigmoid(f_st - f_sc.detach())
    out_to_sc = T.sigmoid(f_to - f_sc.detach())
    out_to_st = T.sigmoid(f_to - f_st.detach())

    loss = T.mul(out_st_sc, lab_sc).mean() \
         + T.mul(out_to_sc, lab_sc).mean() \
         + T.mul(out_to_st, lab_st).mean()

    return loss
```

Listing 1. A PyTorch-like pseudo code for the *Hierarchical Ranking Loss*. Note that confidences of segmentation and labels on different hierarchies are packed into variables 'out' and 'label'.

Ablation of the Stop Gradient. The Stop Gradient operation for each paired loss function converts the problem from a dual optimization problem to a simple regression case. Specifically, the variable detaching from the calculation graph, *i.e.,* Stop Gradient operation, is regarded as a scalar which will not be considered for gradient calculation.

Table A. The F1 scores of the method 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 (w/o SG)	79.3 / **77.2**	74.0 / 73.8	76.2 / 71.8
Multi-Label w/ Hie. Loss (w/ SG)	**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 (w/o SG)	80.1 / **76.9**	75.8 / 74.1	75.5 / 72.3
Multi-Head w/ Hie. Loss (w/ SG)	**80.3** / **76.9**	**76.3** / **74.6**	76.5 / **73.2**

To show the effectiveness of the Stop Gradient (SG) operation in the *Hierarchical Ranking Loss*, we carried out an empirical study to explore the practical performance of the proposed loss with and without the SG operation. As shown in Table A, the overall performances of the loss with SG operation are better than those without SG operation. Besides, in the cross-domain setting, when the SG is not applied, the performance degradation of task *Topic* is the largest compared to other tasks, for instance, 1.3 % and 1.1 % performance drops are observed under the Multi-Label, and Multi-Head protocol, respectively.

C. More Experimental Details and Results

NewsNet holds various hierarchical annotations, enabling the community to explore how to comprehensively represent the complex structure of long-form videos. Thanks to such rich annotations, many video benchmarks can be performed on it. We have conducted a lot of experiments on hierarchical temporal segmentation in the main paper. Here, we complement more experimental details. In addition, we construct baselines on top of state-of-the-art approaches in some video common tasks on our dataset.

C.1. Temporal Video Segmentation

This section provides more experimental details used in both the main paper and the *Supp.*, as well as two more experiments of video temporal segmentation, *i.e.*, modeling the task with conditional random fields (CRF), and *Event* level video temporal segmentation, respectively.

Experimental Details. For the hyperparameters of training the boundary-based model in the experiments, the batch size is set to 16 and the sequence length is set to 40. We use SGD optimizer with 1e-1 initial learning rate and 1e-5 weight decay, and the epoch is set to 300. The dimensions of visual, audio, textual embeddings are 2048, 2048 and 728, respectively.

Cross Entropy *V.S.* CRF. Recently, some approaches [7] convert the temporal segmentation task as a Named Entity

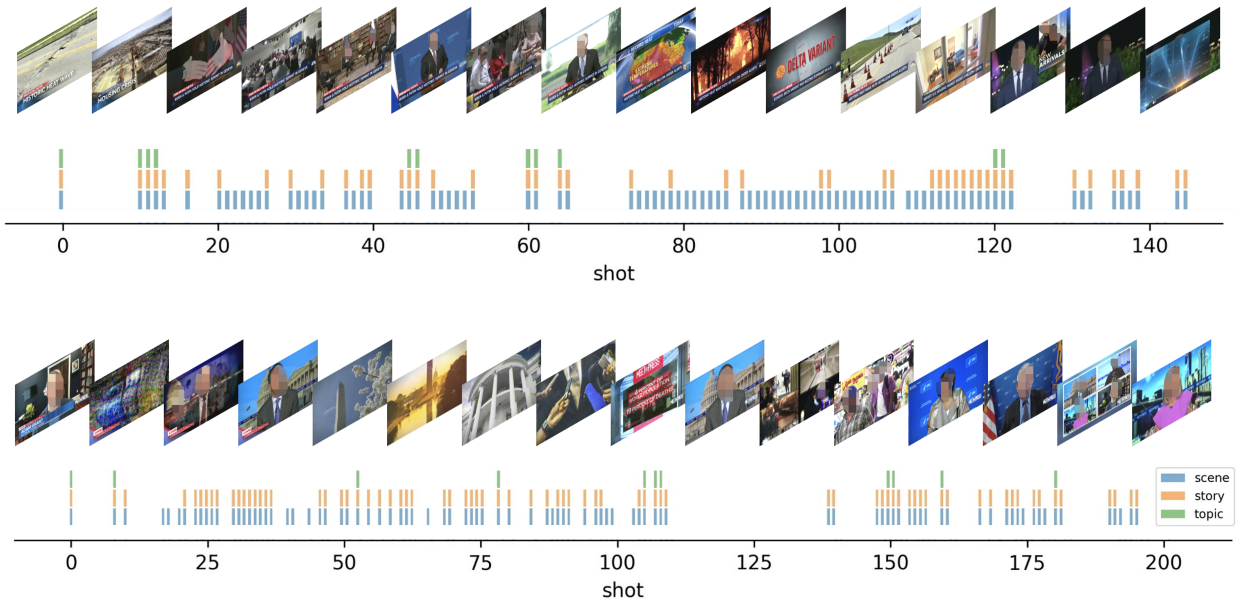

Figure C. **Two annotation examples.** Each colored vertical bar corresponds to the boundary of a specific level, while the distribution and alignment between different levels can be clearly represented on this figure.

Recognition (NER) task, further treating a video clip with the same semantics as an independent entity word. Then a probabilistic model like CRF can be applied to solve and optimize the objective. For a complete discussion of the related methodology, we try to tackle this problem with CRF. More specifically, we first convert the label to the BIES (begin, internal, end and single) way [7] and replace cross entropy Loss with CRF objective [7]. Table. B, lists the results under the in-domain protocol.

Table B. Boundary-free (B.F.) model with Cross Entropy (CE) objective *V.S.* boundary-free (B.F.) model with Conditional Random Fields (CRF) objective in full modality.

Task	Model	F1 score	Precision	Recall
Scene	B.F. w/ CE	78.3	80.9	75.8
	B.F. w/ CRF	**78.4**	80.4	76.5
Story	B.F. w/ CE	**75.4**	74.7	76.2
	B.F. w/ CRF	74.0	75.8	72.3
Topic	B.F. w/ CE	73.2	74.3	72.2
	B.F. w/ CRF	**74.0**	79.3	69.3

Table. B shows that the performance of CRF method has a moderate lead than that of CE method. Yet, we still recommend using CE in the experiment, for the computational efficiency and high compatibility with other approaches, *e.g., Hierarchical Ranking Loss*.

Event Level Temporal Segmentation. In the main paper, due to the need to reduce video redundancy and perfectly align boundaries, we only utilize three hierarchical levels, *i.e., Scene, Story, and Topic*), to discuss the task of hierarchical temporal segmentation. Here, we will complement *Event* level experiments of temporal segmentation. More specifically, we adopt the training protocol used in [4].

Table C. Results of *Event* level temporal segmentation. IM represents ImageNet-1k [1] dataset.

Task	Feature	F1 score	Precision	Recall
Event	ViT [2] w/ IM	34.5	42.2	38.0

Compared to other tasks of temporal segmentation, the result of task *Event* is relatively poor. The possible reasons are: (i) Frame-level input will bring more redundancy and increase difficulties of temporal modeling; (ii) The features of image classification are difficult to provide enough information to identify different people.

C.2. Video Classification

We conduct video classification experiments on the levels of *Topic* and *Story* on the NewsNet. For the backbones, a CNN-based model, *i.e.*, SlowFast [3], and a Transformer-based model, *i.e.*, VideoSwin [8], are selected to tackle the video classification task.

Experimental Details. For each video clip, 16 frames are sampled uniformly for both training and evaluation. All the experiments are trained for 60 epochs with 1e-2 initial learning rate, and a cosine learning rate decay is ap-

plied. We use default model settings of SlowFast [3] and VideoSwin [8]. The resolution of the image is set to 224 × 224, random cropping and horizontal flip are used for augmentation.

Table D. Results of video classification in the visual modality.

Task	Number of Class	Model	Accuracy
Story	10	SlowFast [3]	65.4
		VideoSwin [8]	**72.8**
Topic	6	SlowFast [3]	34.5
		VideoSwin [8]	**42.7**

As shown in Table D, the performances of VideoSwin significantly outperform those of SlowFast for two tasks. Furthermore, all the results on task *Topic* are moderate.

C.3. Video Localization

Experimental Details. We follow the protocol and pipeline of BMN [5] to perform the video localization task on NewsNet. Because the time span variance of *Topic* is too large, we only conduct experiments on *Story*.

Table E. Results of video localization.

Task	Method	AR@50	AR@100	AR@200
Story	BMN [5]	21.4	26.7	32.2

As shown in Table E, BMN [5] achieves 21.4 in terms of AR@50. One possible reason is that the data patterns in news videos can be more complex than those in human action videos. Hence, more multi-modal information may be involved to obtain higher performance on this task in the future.

C.4. Video Highlight Detection

We have introduced how we annotate the highlight news in the Sec. A.2. In this section, we will verify the feasibility of highlight detection on the NewsNet and give some baseline results using the previous state-of-the-art methods.
Highlight Generation. When we have paired NewsFlash clip and Regular News clip, we use shot-level feature similarity to match the shots from the two clips to generate the label of highlight. Thus, we can label the highlight shot in the Regular News clip and treat the highlighted shot as supervision. Finally, the Regular News clips with the annotation of highlight shots are used for the experiments. More specifically, we select three combinations of modality, *i.e.,* visual, textual, and visual+textual, as highlight cues to generate highlight labels.
Experimental Details. We utilize UMT [6] to perform the highlight detection experiment, and the query generator is removed for the absence of query-based textual information on NewsNet. The batch size is set to 4 and we train the model for 200 epochs.

Table F. Results of highlight detection using visual and audio modalities under different highlight cue combinations.

Method	Highlight Cue	mAP
UMT [6]	Visual	43.6
	Textual	25.1
	Visual + Textual	**44.2**

Table F lists the results of highlight detection. As shown in Table F, when both visual and textual modalities are used as the highlight cue, the method achieves the best performance. A potential inspiration is that it is more desirable to discover highlight clips from a multimodal perspective for highly organized videos such as news.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 4

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 4

[3] 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. 4, 5

[4] 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. 4

[5] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. Bmn: Boundary-matching network for temporal action proposal generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3889–3898, 2019. 5

[6] Ye Liu, Siyuan Li, Yang Wu, Chang-Wen Chen, Ying Shan, and Xiaohu Qie. Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3042–3051, 2022. 5

[7] 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. 3, 4

[8] 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. 4, 5

[9] Tomáš Souček and Jakub Lokoč. Transnet v2: an effective deep network architecture for fast shot transition detection. *arXiv preprint arXiv:2008.04838*, 2020. 1

[10] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1