# Collaborative Noisy Label Cleaner: Learning Scene-aware Trailers for Multi-modal Highlight Detection in Movies (Supplementary Material)

Bei Gan Xiujun Shu* Ruizhi Qiao* Haoqian Wu Keyu Chen Hanjun Li Bo Ren
Tencent YouTu Lab
{stylegan, xiujunshu, ruizhiqiao, linuswu, yolochen, hanjunli, timren}@tencent.com


Figure I. Summary of MovieLights.

This supplemental material presents more details of our proposed CLC. Appendix I describes the Multi-modality Sample Cleaning. Appendix II lists some visualization examples and qualitative results in MoiveLights. Finally, we present the broader impact in Appendix III.

## I. Multi-modality Sample Cleaning

The detailed procedure of our MMC is presented in Algorithm 1. The MMC module contains three branches, *i.e.,* multi-modal branch $G^{MM}$, visual branch $G^{UM_v}$, and audio branch $G^{UM_a}$. First, all instances in the two uni-modal branches are fed into the network to obtain the uni-modal losses, *i.e.,* $\mathcal{L}^{UM_v}$ and $\mathcal{L}^{UM_a}$. Next, we select a proportion of instances $N^v$ and $N^a$ that have small training losses in each branch independently. The number of instances is controlled by $\tau$. Then, we combine the selected samples $N^v$ and $N^a$ and take them as clean samples to train the multi-modal branch. Assume that the multi-modal loss is denoted as $\mathcal{L}^{MM}$. Finally, we aggregate all the losses in three branches to update the network parameters. Through joint optimization, the multi-modal branch progressively attains more trustable labels that make learning more robust.

## II. Visualization

As shown in Fig. I, we present a summary to introduce our MovieLights. Fig. II presents the badcase on MovieLights. It incorrectly localized a highlight moment where one of the annotators regards it as a highlight and the other doesn't. Since the subjectivity of the highlight detection, some highlights moment we detect maybe attract only a part of the audience. It reflects the challenge of video highlight detection, and we hope to find better solutions in future work. As seen in Fig. III, the predicted scores detected by CLC and UMT [1] are shown by lines. The results show that in different highlightness clips, the prediction score of UMT [1] has a slight change, while the highlight moments and background scenes can be well distinguished by our CLC. It indicates that our CLC is better at understanding movies and is more discriminative for highlights.

## III. Broader Impact

With the growing number of new publications of movies and the rapid rise of short videos, it is necessary to train automatic movie highlight detection algorithms. This work provides a scene-aware paradigm to learn highlight moments in movies without any manual annotation. Besides, our introduce a framework named Collaborative noisy Label Cleaner (CLC) to learn from these pseudo noisy labels. Finally, the collected dataset MovieLights could foster the further study of movie analysis. The potential negative impact lies in that this dataset may be abused and may cause copyright issues. Hence, to avoid privacy and copyright issues, trailers and movies will be released in the form of extracted features in visual and audio modalities. If actual business data needs to be applied, it should be regulated and consented to by media providers.

*Corresponding author.

**Algorithm 1:** Multi-modality Sample Cleaning

**Input:** Training dataset $D$, a multi-modal branch $G^{MM}$, two uni-modal branch $G^{UM_v}$ and $G^{UM_a}$, clean sample proportion $\tau \in [0, 1]$, iteration $I_{max}$, epoch $E_{max}$

**Output:** Updated branch $G^{MM}$, $G^{UM_v}$ and $G^{UM_a}$

**for** $E = 1, 2, ..., E_{max}$ **do**

  **for** $I = 1, 2, ..., I_{max}$ **do**

    Sample a mini-batch $N$ from the dataset $D$

    Calculate $\mathcal{L}^{UM_v}$ and $\mathcal{L}^{UM_a}$ using the training samples $N$, respectively.

    Filter the noisy labels in each modality

    $N^a = \arg\min_{\tilde{N}^a:|\tilde{N}^a|\geq\tau|N|} \mathcal{L}^{UM_a}(N)$

    $N^v = \arg\min_{\tilde{N}^v:|\tilde{N}^v|\geq\tau|N|} \mathcal{L}^{UM_v}(N)$

    $N'$=$N^v \bigcup N^a$

    Calculate $\mathcal{L}^{MM}$ using the training samples $N'$

    Update $G^{MM}$ by $\mathcal{L}^{MM}$

    Update $G^{UM_a}$ by $\mathcal{L}^{UM_a}$

    Update $G^{UM_v}$ by $\mathcal{L}^{UM_v}$

  **end**

**end**

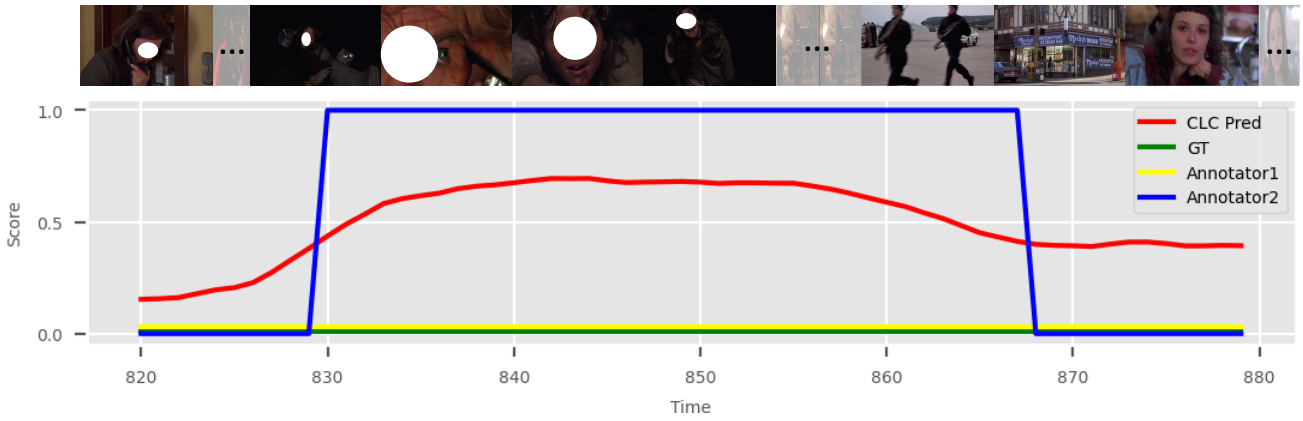


Figure II. Badcase on MoiveLights

## References


[1] 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 *CVPR*, 2022. 1

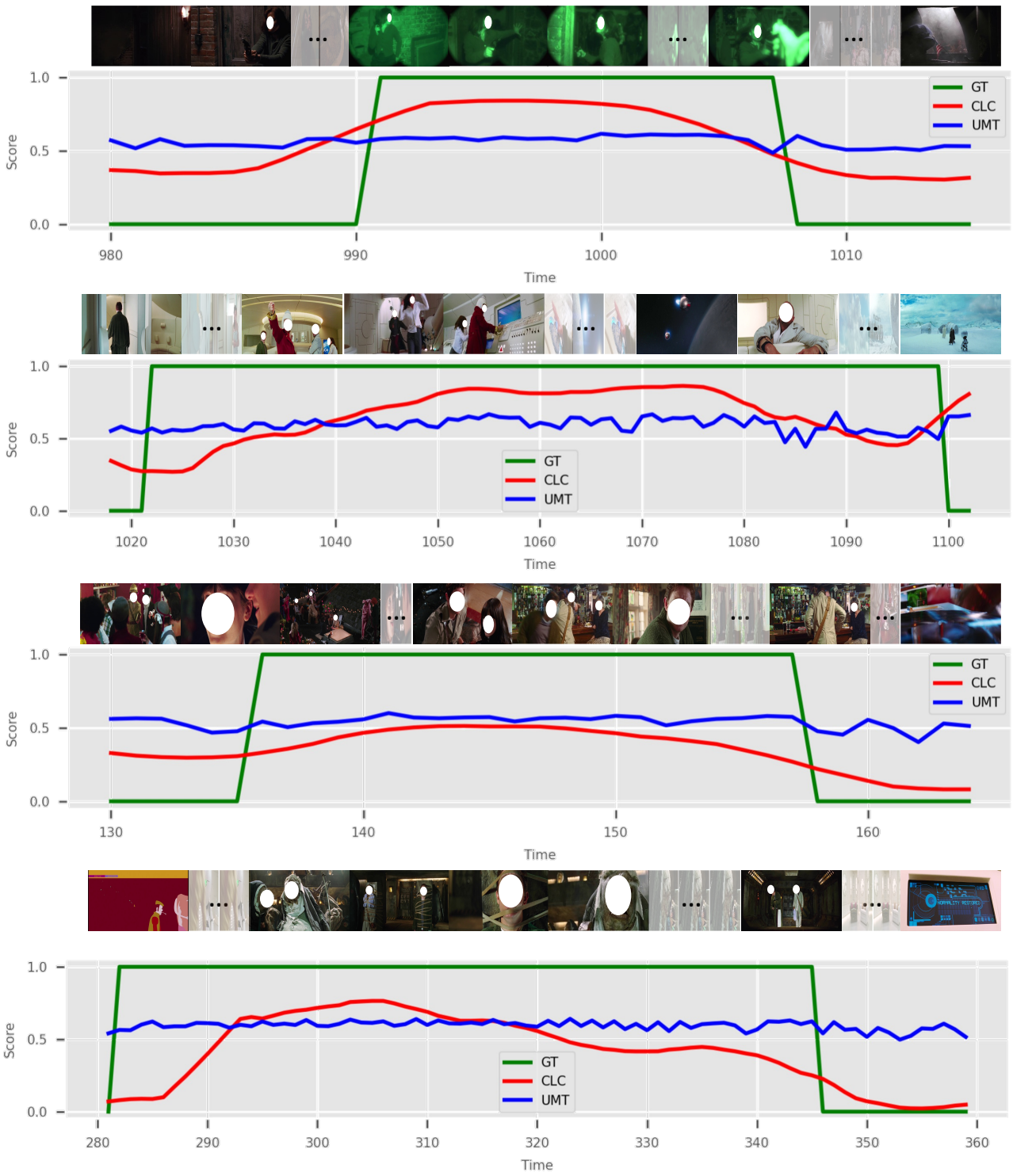


Figure III. Qualitative results. Prediction curve on MoivieLights detected by CLC and UMT.