Collaborative Noisy Label Cleaner: Learning Scene-aware Trailers for Multi-modal Highlight Detection in Movies

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

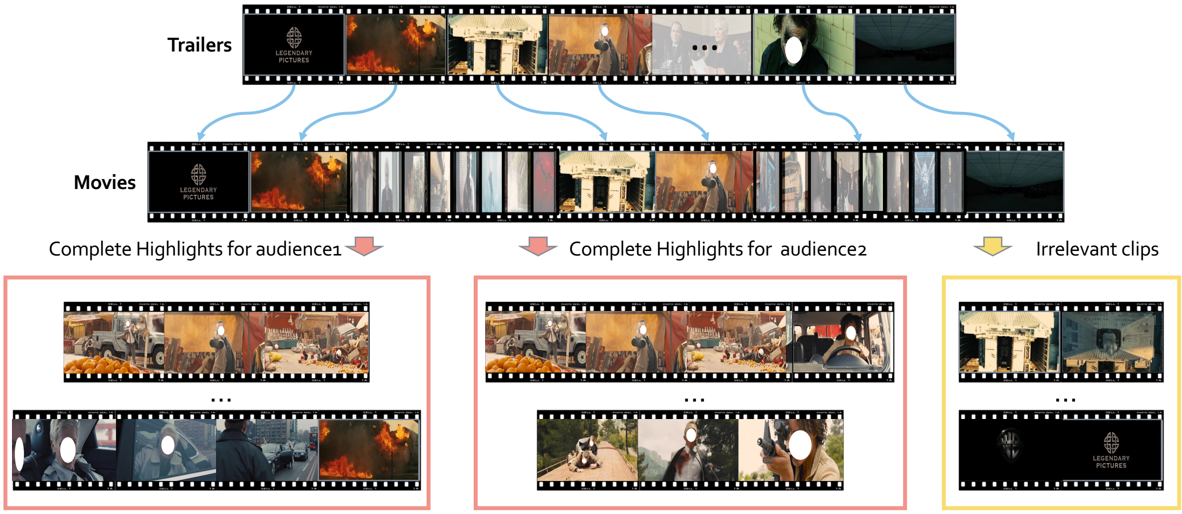
Trailers
Movies
Complete Highlights for audience1
Complete Highlights for audience2
Irrelevant clips

Figure 1. As the preview of the movie, trailers are selected by professionals to grab an audience's attention.However, trailers are usually composed with shots sparsely selected from movies to avoid spoilers, and the audience cannot get complete highlight information.Some trailer clips convey the artistic style of the film only and lack movie storylines, disturbing the audience's impressions.In addition, different audiences may be interested in different styles of clips, which makes it challenging to learn highlights from them.

Abstract

*Movie highlights stand out of the screenplay for efficient browsing and play a crucial role on social media platforms. Based on existing efforts, this work has two observations: (1) For different annotators, labeling highlight has uncertainty, which leads to inaccurate and time-consuming annotations. (2) Besides previous supervised or unsupervised settings, some existing video corpora can be useful, e.g., trailers, but they are often noisy and incomplete to cover the full highlights. In this work, we study a more practical and promising setting, i.e., reformulating highlight detection as "learning with noisy labels". This setting does not require time-consuming manual annotations and can fully utilize existing abundant video corpora. First, based on movie trailers, we leverage scene segmentation to obtain complete shots, which are regarded as noisy labels. Then, we propose a Collaborative noisy Label Cleaner (CLC) framework to learn from noisy highlight moments. CLC consists of two modules: augmented cross-propagation (ACP) and multi-modality cleaning (MMC). The former aims to exploit the closely related audio-visual signals and fuse them to learn unified multi-modal representations. The latter aims to achieve cleaner highlight labels by observing the changes in losses among different modalities. To verify the effectiveness of CLC, we further collect a large-scale highlight dataset named MovieLights. Comprehensive experiments on MovieLights and YouTube Highlights datasets demonstrate the effectiveness of our approach. Code has been made available at:*https://github.com/TencentYoutuResearch/HighlightDetection-CLC.

1. Introduction

With the growing number of new publications of movies in theaters and streaming media, audiences become even

*Corresponding author.

harder to choose their favorite one to enjoy for the next two hours. An effective solution is to watch the movie trailers before choosing the right movie. This is because trailers are generally carefully edited by filmmakers and contain the most prominent clips from the original movies. As a condensed version of full-length movies, trailers are elaborately made with highlight moments to impress the audiences. Consequently, they are high potential in serving as supervision sources to train automatic video highlight detection algorithms and facilitating the mass production of derivative works for video creators in online video platforms, e.g., YouTube and TikTok.

Existing video highlight detection (VHD) approaches are generally trained with annotated key moments of long-form videos. However, they are not suitable to tackle the movie highlight detection task by directly learning from trailers. The edited shots in trailers are not equivalent to ground-truth highlight annotations in movies. Although a previous work [43] leverages the officially-released trailers as the weak supervision to train a highlight detector, the highlighted ness of trailer shots is extremely noisy and varies with the preference of audiences, as shown in Fig. 1. On one hand, trailers tend to be purposefully edited to avoid spoilers, thus missing key moments of the storylines. On the other hand, some less important moments in the original movies are over-emphasized in the trailers because of some artistic or commercial factors. The subjective nature of trailer shots makes them noisy for the VHD task, which is ignored by existing VHD approaches.

To alleviate the issue, we reformulate the highlight detection task as "learning with noisy labels". Specifically, we first leverage a scene-segmentation model to obtain the movie scene boundaries. The clips containing trailers and clips from the same scenes as the trailers provide more complete storylines. They have a higher probability of being highlight moments but still contain some noisy moments. Subsequently, we introduce a framework named Collaborative noisy Label Cleaner (CLC) to learn from these pseudo-noisy labels. The framework firstly enhances the modality perceptual consistency via the augmented cross-propagation (ACP) module, which exploits closely related audio-visual signals during training. In addition, a multi-modality cleaning (MMC) mechanism is designed to filter out noisy and incomplete labels.

To support this study and facilitate benchmarking existing methods in this direction, we construct MovieLights, a Movie Highlight Detection Dataset. MovieLights contains 174 movies and the highlight moments are all from officially released trailers. The total length of these videos is over 370 hours. We conduct extensive experiments on MovieLights, in which our CLC exhibits promising results. We also demonstrate that our proposed CLC achieves significant performance-boosting over the state-of-the-art on the public VHD benchmarks.

In summary, our major contributions are as follows:

- We introduce a scene-aware paradigm to learn highlight moments in movies without any manual annotation. To the best of our knowledge, this is the first time that highlights detection is regarded as learning with noisy labels.
- We present an augmented cross-propagation to capture the interactions across modalities and a consistency loss to maximize the agreement between the different modalities.
- We incorporate a multi-modality noisy label cleaner to tackle label noise, which further improves the robustness of networks to annotation noise.
- Experiments on movie datasets and benchmark datasets validate the effectiveness of our framework.

2. Related Works

Video Highlight Detection. This task aims to identifying the interesting moments from untrimmed videos. In recent years, the videos studied for this task extend from domain-specific sport videos [41] to general videos such as social media videos [37], news [35], first-person videos [53] and vlog [25] . Most of previous works [14, 21, 53] interpret the video highlight detection task as a segment-level ranking problem. They compare pairwise segments from same domain video in order to learn a model that assigns highlight scores to these segments where the highlight segments receive higher scores than the non-highlight segments. MINI-Net [16] proposes to cast highlight detection as multiple instance ranking network learning. SL-Module [50] explores the highlight detection problem through Unsupervised Domain Adaptation (UDA) [32]. UMT [29] integrates highlight detection and moment retrieval into a unified framework and conduct joint optimization. PLD [45] models the video highlight detection into a pixel-level distinction estimation task. In this work, we regard highlight detection as learning with noisy labels. Joint-VA [2] also considers video highlight detection from the perspective of noise. However, it focuses on noise in features, such as videos having noisy audio when the microphone constantly has water splashing against it. We focus on specific annotation noise in video highlight detection.

Studies on Movies and Trailers. Studies on movies and trailers have received increased attention in research. [18] introduces a comprehensive dataset for movie understanding. [4, 42] try to model the relationships among the movie characters. [7, 34, 47] focus on breaking the storylines of movies into semantically cohesive parts. Besides the studies on movies, efforts have been made to develop trailer understanding. [57] presents a movie summarization system and composes movie summaries in terms of user experience evaluation. [9] designs a movie trailer dataset for the evaluation of video-based recommender systems. [19] is the first approach that bridges trailers and movies and allows knowledge learned from trailers to be transferred to full movie analysis. In [43], the visual module and the temporal analysis module are respectively trained on trailers and movies. Because of the inaccessibility of public trailer-related benchmarks, we construct a new dataset (MovieLights) to detect the highlight moments in movies.

Learning with Noisy Labels. Learning with noisy labels has been a long-standing problem in computer vision. There are three kinds of approaches to this problem. One of the most common strategies for tackling label noise is to capture the transition probabilities between noisy labels and clean labels [31,38,48,49,52,56]. Another solution is to design robust loss functions for model training against noisy labels [8, 26, 28, 30, 44, 55]. A popular method is to design a mechanism to select clean samples or give lower weight for noise samples in the training set to reduce impact of noise [15, 17, 20, 39, 46, 51]. In this paper, we attempt to solve the problem by exploiting multi-modalities nature of movies.

3. Movie Highlight Dataset

In this paper, we aim to detect the highlight moments in movies by learning from easily accessible trailers as the noisy supervision. However, the existing movie and trailer-related benchmarks [3, 9, 18] lack sufficient functions for this task, such as the absence of full-length movies and ground-truth highlight annotations. Huang et al. [19] propose to respectively learn visual representations from trailers and temporal structure from full-length movies in their constructed Large-Scale Movie and Trailer Dataset (LSMTD). Wang [43] constructs a Trailer Moment Detection Dataset (TMDD) for detecting trailer moments from full-length movies without explicit human annotation. Both LSMTD and TMDD are not publicly available, while TMDD only contains three movie genres.

The inaccessibility of public benchmarks motivates us to construct a new dataset, named Movie Highlight Detection Dataset (MovieLights). In particular, we purchase a set of movies from commercial channels and collect their corresponding trailers from streaming platforms such as YouTube, covering at least 25 genres to ensure content diversity. The movies and trailers are then prepossessed by shot segmentation [36] and scene segmentation [47], respectively. The resulting shots are a series of consecutive frames taken by the camera until a physical interruption, and the scenes are consecutive shots that share a semantically related theme.

As seen in Fig. 2, to build the ground truth, we conduct Faiss [22] to obtain visual similarity matching between trailer frames and movie frames. We locate trailer moments in the movie and align them with the movie shots as annotation references. For the testing set, we collect 2 sets of moments for each movie from different workers, and these moments are annotated by different annotators independently. To ensure the consistency of results from different annotations, during the annotation procedure, all highlight moments must be related to the annotation references. Though all selected shots are relevant to the trailers, as highlight moments can be subjective, they may still vary in their saliency and time span. We calculate the intersection between every pair of moments annotated as the ground truth. However, the vast diversity of movie storylines makes the annotation challenging as it is time-consuming and requires annotators to be familiar with the movie. To collect a large amount of training data efficiently, we introduce a scene-aware paradigm to obtain the highlight moments label without any manual annotation. Specifically, we expand the shot-level annotation references to the scene span as positive samples automatically. It will capture the complete scene context of the trailer shot with movie storylines. Since the trailer shots may contain some less important moments, the acquired highlight labels are still noisy.

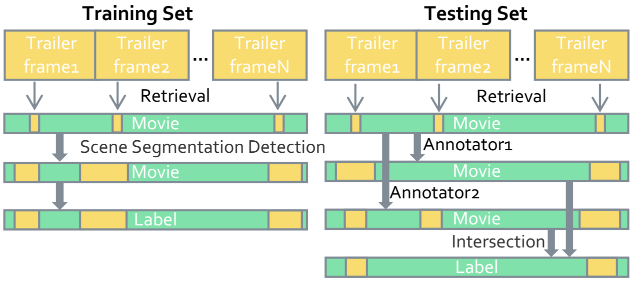

Figure 2. The labeling process of MovieLights. For the training set, we introduce a scene-aware paradigm to obtain labels automatically. For the testing set, we collect 2 sets of labels from different annotators.

Table 1. The basic statistics of MovieLights.

	Train	Test
Movie Number	144	30
Avg Durations per Movie	2.19h	2.14h
Avg Shot Number per Movie	1852	1940
Avg Scene Number per Movie	207	193
Annotator1 Positive sample Proportion	-	0.27
Annotator2 Positive sample Proportion	-	0.30
Positive sample proportion	0.35	0.21

As seen in Tab. 1, MovieLights contains 174 movies in full length with their official trailers and it is split into a training set with 144 movies, and a testing set with 30 movies. The content diversity is ensured by the rich domain informations (more than 25 genres) and abundant segments

($325k$ shots and $36k$ scenes). Most movies in our dataset are between 90 to 150 minutes, and the length of the annotated moments varies from tens of seconds to several minutes. As the acquired positive highlight moments take up 35% in the training set while the annotated true positives take up 21% in the testing set, the difference tells the obvious existence of label noise.

We plan to release the dataset publicly to promote further study of movie analysis. Due to copyright issues, trailers and movies will be released in the form of extracted features in visual and audio modalities.

4. Approach

4.1. Overview

The overall architecture of our Collaborative noisy Label Cleaner (CLC) framework is illustrated in Fig. 3, which includes three modules: feature extraction, augmented cross-propagation (ACP), and multi-modal cleaning (MMC).

In our framework, both the visual and audio modalities are utilized. For feature extraction and encoding, we first split the video V into T shots. We characterize the i^{th} shots by two vectors, *i.e.,* $\mathbf{v}_i$ for the visual features, and $\mathbf{\hat{a}_i}$ for the audio features, where $i = 1, 2, ..., T$. These features are extracted using pre-trained visual [11] and audio feature extractors [24]. The parameters of the two extractors are frozen during training.

The ACP and MMC are the core components of our framework. Since the visual-audio signals in videos are closely related but do not always contribute to highlight detection, the ACP module exploits the relationship via uni-modal and cross-modal interactions. Then, it learns unified multi-modal representations for highlight detection. As the obtained highlight moments after scene segmentation are noisy, the MMC module firstly observes the changes in uni-modal losses, then filters the noisy labels and utilizes the clean ones for multi-modal supervised learning. Next, we will introduce details for each component.

4.2. Augmented Cross-Propagation

Cross-Propagation. To predict the highlights, the model needs to understand the storylines of the movie. Meanwhile, visual and audio inputs do not always contribute to accurate prediction. Therefore, temporal modeling and modality interaction are the keys to achieving successful highlight detection. To achieve this goal, we design our ACP module to fuse the visual-audio modalities. It involves three steps in total.

First, in order to align the multi-modal features, we introduce h which is fully-connected (FC) layer with ReLU to the $\mathbf{\hat{a}_i}$ such that it has same dimension as the $\mathbf{v}_i$.

$$\mathbf{a}_i = h(\mathbf{\hat{a}_i}). \tag{1}$$

Movies are composed of consecutive audio-visual clips. Therefore, to measure the highlighted ness of a given shot, one must consider the relationship of the shot with its adjacent shots. The self-attention mechanism has shown effectiveness in capturing the long-term dependencies in previous works. We leverage self-attention to capture the temporal relationship for $\mathbf{v}_i$ and $\mathbf{a}_i$ via Eq. 2 and Eq. 3, respectively.

$$\mathbf{v}_i^s = softmax\left(\frac{(\mathbf{v}_i W_1^v)(\mathbf{v}W_2^v)^\top}{\sqrt{d}}\right)(\mathbf{v}W_3^v), \tag{2}$$

$$\mathbf{a}_i^s = softmax\left(\frac{(\mathbf{a}_i W_1^a)(\mathbf{a}W_2^a)^\top}{\sqrt{d}}\right)(\mathbf{a}W_3^a), \tag{3}$$

where $\mathbf{v} = [\mathbf{v}_1; \mathbf{v}_2; ...; \mathbf{v}_T]$ and $\mathbf{a} = [\mathbf{a}_1; \mathbf{a}_2; ...; \mathbf{a}_T]$; the scaling factor d is equal to the visual/audio feature dimension and $(*)^\top$ denotes the transpose operator; W^v and W^a are learnable matrices of two modalities, which are implemented by a linear layer. Uni-modal self-attention can well capture uni-modal temporal contexts and enhance clip features within the same modality.

Despite the above self-attention capturing the clip interactions within the uni-modality, it is critical to capture the interactions across modalities. To capture semantic associations based on multi-modal signals, we introduce cross-attention to update the features of each modality.

$$\mathbf{v}_i^c = softmax\left(\frac{(\mathbf{v}_i W_4^v)(\mathbf{a}W_4^a)^\top}{\sqrt{d}}\right)(\mathbf{a}W_5^a), \tag{4}$$

$$\mathbf{a}_i^c = softmax\left(\frac{(\mathbf{a}_i W_6^a)(\mathbf{v}W_5^v)^\top}{\sqrt{d}}\right)(\mathbf{v}W_6^v), \tag{5}$$

Through cross-attention, the information from two modalities are connected. However, considering audio-visual temporal asynchrony, it is necessary to select effective information from multi-modality. We augment the relevant positive connections and dampen the irrelevant connections. The strength of these connections is measured by the cross-correlation matrix, computed by,

$$\mathbf{c}^v = ReLU\left(\frac{\mathbf{v}\mathbf{a}^\top}{\sqrt{d}}\right), \mathbf{c}^a = ReLU\left(\frac{\mathbf{a}\mathbf{v}^\top}{\sqrt{d}}\right). \tag{6}$$

For each modality, the cross-correlation matrix $\mathbf{c}$ is used to re-weight the cross-attention features. Finally, we obtain updated visual features $\mathbf{\bar{v}_i}$ and audio features $\mathbf{\bar{a}_i}$ by fusion of the original features, enhanced uni-modal features, and cross-modal features.

$$\mathbf{\bar{v}_i} = f(\mathbf{v}_i, \mathbf{v}_i^s, (\sum_{j=1}^{T} \mathbf{c}_{ij}^v) * \mathbf{v}_i^c), \tag{7}$$

$$\mathbf{\bar{a}_i} = f(\mathbf{a}_i, \mathbf{a}_i^s, (\sum_{j=1}^{T} \mathbf{c}_{ij}^a) * \mathbf{a}_i^c), \tag{8}$$

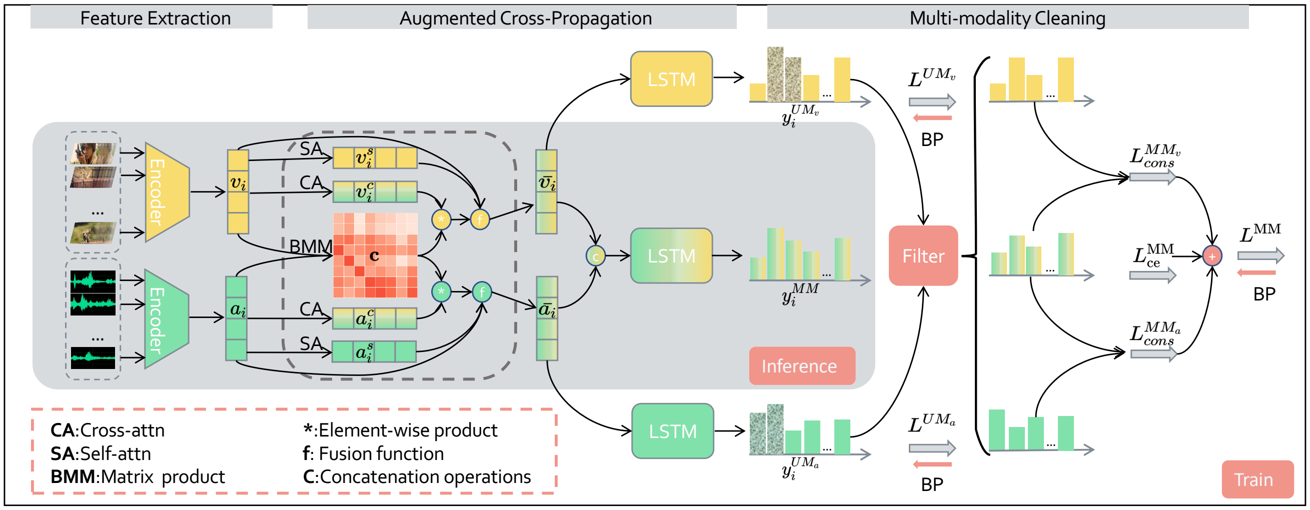

Figure 3. Overview of the proposed CLC. It includes three modules: feature extraction, augmented cross-propagation (ACP), and multi-modal cleaning (MMC). The visual and audio modalities of the input video are represented as vectors by the feature extraction module. Then the features are augmented by ACP module to capture semantic associations acorss modalities. MMC is used to filter outs noisy and incomplete labeling with additional uni-modal branches. During inference, we remove the uni-modal branches and only rely on the prediction of multi-modal branch. More details of the CLC are shown in Sec. 4.1

where f is the fusion function consisting of FC layers and ReLU to further project the features.

Clearly, the cross-correlation matrix will assign large weights to clips that are relevant to the other modality. In addition, the ReLU activation in Eq. 6 cuts off connections with negative similarity values and only relevant positive connections would be preserved. By the above operations, the original feature will be infused with richer information.

Consistency Loss. Multi-modal inputs help to comprehensively learn by integrating different aspects and boosting model performance. However, they are not fully exploited because some modality-specific features may be weakened in the fusion even when the multi-modal model outperforms its uni-modal counterpart. In this work, we provide parallel branches for each modality separately to obtain uni-modal prediction during training. As seen in Fig. 3, all branches share the same clip feature extracted from ACP and the feature is fed into different branches independently. In each branch, we develop a temporal model G to obtain its prediction score $\mathbf{y}_i$ of the i^{th} clip being a highlight as follows:

$$\mathbf{y}_i^{MM} = G(\bar{\mathbf{v}}_\mathbf{i}, \bar{\mathbf{a}}_\mathbf{i}), \mathbf{y}_i^{UM_v} = G(\bar{\mathbf{v}}_\mathbf{i}), \mathbf{y}_i^{UM_a} = G(\bar{\mathbf{a}}_\mathbf{i}). \quad (9)$$

where MM is multi-modal branch; UM_v is visual branch and UM_a is audio branch.

However, the gap between the different modalities would bring instability in the joint optimization process. We employ the auxiliary consistency loss to guarantee consistency between the different modalities. Given the multi-modal features of a clip, they are consistent if they share the same prediction. Specifically, the consistency loss is defined as the cross-entropy between the multi-modal prediction probability $\mathbf{y}_i^{MM}$ and uni-modal prediction probability $\mathbf{y}_i^{UM}$:

$$\mathcal{L}_{\text{cons}} = -\left(\sum_{i=1}^{N} \mathbf{y}_i^{UM_v} \log \mathbf{y}_i^{MM} + \sum_{i=1}^{N} \mathbf{y}_i^{UM_a} \log \mathbf{y}_i^{MM}\right), \quad (10)$$

where N is the number of samples in a batch.

This consistency loss not only implicitly enhances uni-modal information, but also explicitly guides the multi-modal branch to robuster supervision. The added uni-modal branches are only utilized in the training phase and are disabled during the inference stage.

4.3. Multi-modal Cleaning

Multi-modality Sample Cleaning. Annotation noise is inevitable in VHD due to subjectivity depending on the users and annotators. In this paper, we argue that highlight detection should be regarded as Learning with Noisy Labels. To alleviate the performance drop caused by noisy labels, we adopt a multi-modality collaborative cleaning to adaptively filter noisy samples with noisy modality information.

Firstly, we maintain multiple outputs simultaneously which are predicted by different branches. The uni-modal branches independently select clean samples based on the low-loss criterion in which instances with lower losses are treated clean samples. Contrary to existing noisy sample selection methods [12, 54], which directly discard high-loss samples, we keep all samples to train the uni-modal branches. Then, we update the multi-modal branch using only clean samples selected by both uni-modal branches in the back-propagation. The samples for the multi-modal branch training are selected dynamically while all samples participate in the uni-mode branches training. In this way, MMC obtains information on all samples to avoid the model defecting to favoring easy samples. More details of MMC are presented in the supplementary material.

Each branch has its own loss function. For the uni-modal branch, we employ the cross-entropy loss with all samples as follows:

$$\mathcal{L}^{UM_v} = \mathcal{L}_{\text{ce}}^{UM_v} = -\sum_{i=1}^{N} \mathbf{g}_i \log \mathbf{y}_i^{UM_v}, \tag{11}$$

$$\mathcal{L}^{UM_a} = \mathcal{L}_{\text{ce}}^{UM_a} = -\sum_{i=1}^{N} \mathbf{g}_i \log \mathbf{y}_i^{UM_a}. \tag{12}$$

For the multi-modal branch, its cross-entropy loss and consistency loss are updated with a re-weighting scheme with clean samples.

$$\mathcal{L}_{\text{ce}}^{\text{MM}} = -\sum_{i=1}^{N'} \mathbf{g}_i \log \mathbf{y}_i^{MM}, \tag{13}$$

$$\mathcal{L}_{\text{cons}}^{MM} = -\left(\sum_{i=1}^{N'} \mathbf{y}_i^{UM_v} \log \mathbf{y}_i^{MM} + \sum_{i=1}^{N'} \mathbf{y}_i^{UM_a} \log \mathbf{y}_i^{MM}\right), \tag{14}$$

$$\mathcal{L}^{\text{MM}} = \mathcal{L}_{\text{ce}}^{\text{MM}} + \beta \mathcal{L}_{\text{cons}}^{MM}, \tag{15}$$

where $\mathbf{g}_i$ and $\mathbf{y}_i^*$ denote the ground-truth and predicted probability of the i^{th} clip, respectively; N is the number of samples in a batch and N' is the number of samples seleted by MMC; and β are designed to balance different loss terms.

Post processing. In experiments we observe that noise makes jitter prediction curves along the temporal dimension, which may cause discontinuous thresholds for highlight selection. Therefore, we apply a median filter to smooth the prediction curves. Supposing $\mathbf{y} = [\mathbf{y}_1; \mathbf{y}_2; ...; \mathbf{y}_T]$ is the original curve predicted by the CLC, the smoothed curve $\mathbf{s} = [\mathbf{s}_1; \mathbf{s}_2; ...; \mathbf{s}_T]$ is given by:

$$\mathbf{s}_i = \begin{cases} \text{Med}\left(\mathbf{y}_{i-k}, \mathbf{y}_{i+k}\right), & k < i \leq T - k \\ \mathbf{y}_i, & \text{otherwise} \end{cases} \tag{16}$$

where k is the window size, and "Med" denotes the median filter.

5. Experiment

5.1. Datasets and Experimental Settings

Datasets. We evaluate our CLC on the constructed dataset MovieLights and public YouTube Highlights dataset [37]. MovieLights is split into training and testing sets, each containing 144 and 30 movies respectively.

We split movies into shots. The movie features are represented at shot-level. We use the middle frame of each shot to extract its visual feature with ViT [10] pre-trained by CLIP [33]. We align timestamps of audio clips with the visual shots, and sample the audio clip of each with $16K$ Hz sampling rate and 512 windowed signal length. The resulted shot-level audio features are obtained with the PANN audio network [24] pretrained on AudioSet [13].

The YouTube Highlights contains six distinct categories with a total 422 videos currently available. Following the practice of prior efforts, we train a highlight detector for each category. YouTube Highlights provides two annotations: Harvested Highlight and Mturk Highlight. In the Harvested annotation, the match label specifies if each clip is matched in the edited video, where 1 denotes matched, -1 denotes unmatched and 0 denotes the borderline cases. In the Mturk annotation, the highlight labels are marked by multiple turkers of different styles, making it noisier than the Harvested annotation.

On YouTube Highlights, we use the same protocol and data preprocessing as [29]. It obtain clip-level visual features and optical flow features using I3D [6] pre-trained on Kinetics400 [23]. It use a PANN audio network [24] pre-trained on AudioSet [13] to obtain audio features that align with the visual clips. Frame-level features are average-pooled within each clip for both audio and visual features to generate a clip-level feature. Since each feature vector spans 32 consecutive frames, we follow [29] and consider the feature vector corresponded to a clip if their overlap is more than 50%.

Benchmarks. To better inspect the robustness of our CLC against noisy labels, we also apply label perturbations in training set of YouTube Highlights while keeping the ground-truths in the testing set unchanged. YouTube Highlights has two benchmarks for comparison. 1) **Harvested with matched:** we regard the clips labeled with matched as the highlighted clips and this is the same setting as in previous works [29, 50]. 2) **Harvested with matched and borderline:** clips labeled with matched and borderline are treated as the highlight moments. In this benchmark, the training set contains some highlighted clips with weak confidence. We select the clips whose mturk-label is over score 1 as the highlighted clips, which means that at least one turker selects the clip as a highlight. There are labeled by different types of annotators between Mturk and the test set and bring greater noise. We select the clips whose mturk-label is over score 1 as the highlighted clips, which means that at least one turker selects the clip as a highlight. Due the labeling gap between the Mturk annotation in the training set and the clean testing set, this benchmark is even noisier.

Baselines. We introduce CLC$-$, the degenerated version of CLC, as a baseline. Similar to CLC, CLC$-$ is a Bi-LSTM-based model and takes the temporal sequence of shot features as input but lacks the modules of augmented cross-propagation and multi-modality noisy label cleaner.

Table 2. Results on MovieLights.

Methods	Modality	mAP
GIFs [14]	V	25.48
SL-Module [50]	V	32.34
SL-Module [50]	VA	34.27
UMT [29]	VA	38.7
CLC−	VA	39.65
CLC− w/ SCE [44]	VA	39.83
CLC− w/ LS [40]	VA	40.49
CLC	VA	**43.88**

Table 3. Ablation results of MoiveLighgts.

MMSC	CP	CL	PP	mAP
×	×	×	×	39.65
✓	×	×	×	41.69
✓	✓	×	×	42.79
✓	✓	✓	×	43.22
✓	✓	✓	✓	**43.88**

Table 4. Results on YouTube Highlights.

Methods	dog	gym.	park.	ska.	ski.	surf.	Avg.
GIFs [14]	30.8	33.5	54	55.4	32.8	54.1	46.4
LSVM [37]	60.0	41.0	61.0	62.0	36.0	61.0	53.6
HighlightMe [5]	63	73	72	64	52	62	64
MINI-Net [16]	58.2	61.7	70.2	72.2	58.7	60.1	64.4
CHD [1]	60.6	71.1	74.2	49.8	68.2	68.5	65.4
Trail [43]	63.3	82.5	62.3	52.9	74.5	79.3	69.1
SL-Module [50]	70.8	53.2	77.2	72.5	66.1	76.2	69.3
Joint-VA [2]	64.5	71.9	80.8	62	73.2	78.3	71.8
PLD [45]	74.9	70.2	77.9	57.5	70.7	79	73
CO-AV [27]	60.9	66	89	74.1	69	81.1	74.7
UMT [29]	65.9	75.2	81.6	71.8	72.3	82.7	74.9
CLC(ours)	70.5	79.4	83.9	83.5	79.5	83.6	**80.1**

Evaluation Metric. We adopt mean Average Precision (mAP) as the evaluation metric for MovieLights and YouTube Highlights. Considering that a highlighted moment in one video is not necessarily more interesting than non-highlight moments in other videos, we evaluate on each test video independently and report the averaged results.

Implementation Details. On the MovieLights, we train our model using SGD, with a learning rate of 0.01. We train for 50 epochs. Before the cross-attention modules, we project each modality into a vector of 512 dimension. The key, query, and value vectors all share the same dimension. Weight β in Eq. 15 is empirically set to 0.1 and window size k in Eq. 16 is set to 9.

5.2. Results on MovieLights

On MovieLights, we train our model with the noisy pseudo labels. To compare with previous state-of-the-art highlight detection works, we train UMT [29] and SL-Module [50] using the same protocol and data preprocessing as in CLC. We also compare with Video2GIF [14] using its off-the-shelf tool[1]. The upper part of Tab. 2 illustrates the significant performance gain of CLC.

To demonstrate the advantages of CLC in learning with noisy labels, we make comparisons with two main-stream label noise approaches: Label Smoothing [40] and SCE loss [44]. Specifically, we insert Label Smoothing or SCE into our CLC− framework to create two baseline VHD methods to tackle label noise. The bottom part of Tab. 2 shows that CLC outperforms the two baseline methods by a notable margin, indicating that our augmented cross-propagation and multi-modality noisy label cleaner are more effective than vanilla label noise approaches in VHD tasks.

5.3. Ablation Study

In this experiment, we analyze the impact of each module. The results are summaried in Tab. 3.

Multi-modality Sample Cleaning. We first inspect the impact of the multi-modality sample filter module because we are primary concerned with learning with noisy labels in VHD. The 2% performance gain from the module over baseline shows the importance of filtering noisy sample. Based on the this observation, the subsequent ablation experiments are conducted under the setting of noise filtering.

Cross-Propagation. We then evaluate the impact of the feature augmentation module. Compared to naive feature concatenation, the models with augmented features show superior performance. The results validate that feature augmentation module can better explore the complementary information from different modalities and suppress the mutual disturbance of desynchronized uni-modal information.

Consistency Loss. We examine the contribution of the consistency loss. The results demonstrate that the employment of the auxiliary multi-modal constraint further increases the model robustness. This consistency loss not only implicitly enhances uni-modal information, but also explicitly guides the multi-modal branch to better learning.

Post Processing. We compare the highlight prediction curves with and without the post processing median filter in Fig. 4(a,b), which prevents disruptive prediction variation and improves the overall performance.

5.4. Results on YouTube Highlights

We conduct experiments on the public video highlight detection benchmarks YouTube Highlights, including six domain video datasets, to verify the generalization ability of CLC.

The setting of **Harvested Highlight with matched** is consistent with the previous work [29]. As shown in Tab. 4, CLC achieves state-of-the-art performance on the YouTube

[1] https://github.com/gyglim/video2gif_code

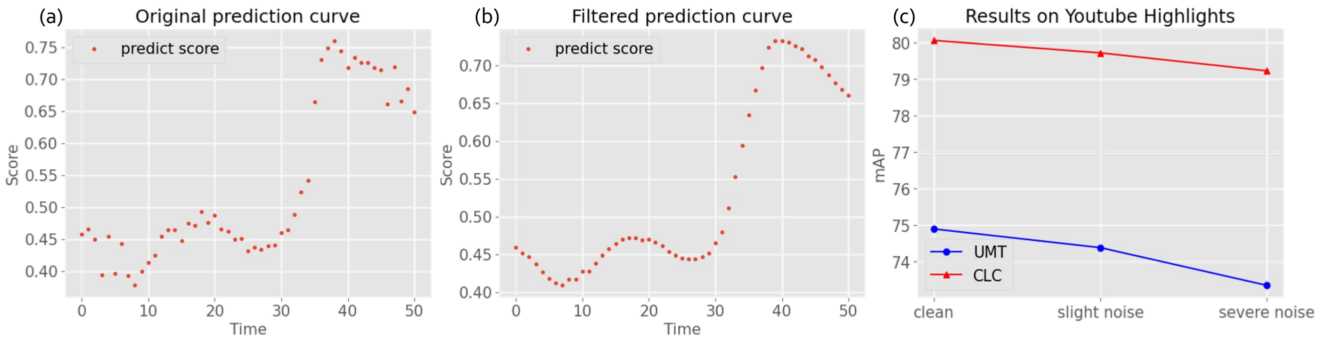

Figure 4. (a, b) Comparison of original prediction curve with filtered prediction curve. (c) Results on YouTube Highlights with noisy label.

Table 5. Results on YouTube Highlights with Noisy Label.

Annotation	Noise	Methods	dog	gym.	park.	ska.	ski.	surf.	Avg.
Harvested matched	clean	UMT [29]	65.90	75.20	81.60	71.80	72.30	82.70	74.90
Harvested matched	clean	CLC	70.51	79.43	83.85	83.51	79.46	83.56	80.05 (↑ 5.15)
Harvested borderline	slight noise	UMT [29]	65.93	74.31	81.58	71.84	70.24	82.46	74.39
Harvested borderline	slight noise	CLC	69.41	80.73	78.50	85.36	81.11	83.16	79.71 (↑ 5.32)
Mturk	severe noise	UMT [29]	63.78	76.16	75.02	73.62	69.99	81.59	73.36
Mturk	severe noise	CLC	66.92	80.44	85.92	82.33	78.05	81.72	79.22 (↑ 5.86)

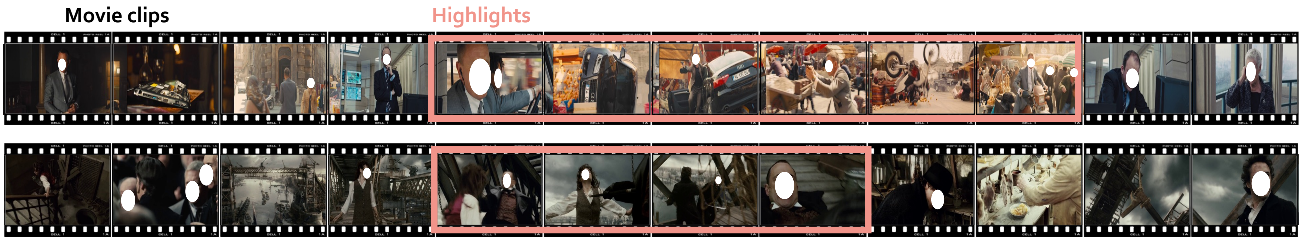

Figure 5. The highlight moments selected by our CLC from Skyfall and Sherlock Holmes. Top: Bond gives chase to a professional hitman by car to find a classified hard drive, and then a firefight erupted in the market. Bottom: Holmes and Blackwood are facing off, and then Blackwood reaches for a weapon to kill Holmes, but accidentally trips off a scaffolding and falls to his death.

Highlights, outperforming the existing multi-modal highlight detection methods in the average metric across all categories. Specifically, CLC achieves best performance in three out of the six categories, while maintaining reasonably competitive performance in the other three categories. These results support our claim that highlight detection should be regarded as learning with noisy labels.

To quantify how CLC is robust to different levels of label noise, we inspect the settings of **Harvested Highlight with matched and borderline** and **Mturk Highlight**, which are perturbed by varying degrees of label noise in YouTube Highlights.

Tab. 5 exhibits the performances of CLC and UMT [29] at different noise levels. As the noise level increases, the VHD task becomes more difficult, but the performance superiority of our CLC over UMT becomes even more obvious. It is illustrated in Fig. 4(c) that compared with the most recent state-of-the-art UMT, our model can achieve better performance even when disturbed by severe label noise.

5.5. Visualization

As shown in Fig. 5, we present some visualization examples of the detected highlight clips in MovieLights by CLC. The examples clearly shows that the prediction of CLC is in accordance with user expectation. We will provide more examples in the supplementary material.

6. Conclusion

In this study, we present Collaborative noisy Label Cleaner (CLC), a novel framework to handle noisy labels in video highlight detection. We make use of the augmented cross-propagation module to better enhance network robustness and multi-modality cleaning to achieve cleaner highlight labels by observing the loss changes of different modalities. We demonstrate the state-of-the-art performance of our method with extensive experiments on MovieLights and YouTube Highlights datasets. In future work, we are interested in extending the proposed mechanisms to other video-understanding tasks such as scene segmentation and video temporal grounding.

References

[1] Taivanbat Badamdorj, Mrigank Rochan, Yang Wang, and Li Cheng. Contrastive learning for unsupervised video highlight detection. In *CVPR*, 2022. 7

[2] Taivanbat Badamdorj, Mrigank Rochan, Yang Wang, Jiaze Wang, and Li Cheng. Joint visual and audio learning for video highlight detection. In *ICCV*, 2021. 2, 7

[3] Max Bain, Arsha Nagrani, Andrew Brown, and Andrew Zisserman. Condensed movies: Story based retrieval with contextual embeddings. In *ACCV*, 2020. 3

[4] David Bamman, Brendan O'Connor Noah, and A. Smith. Learning latent personas of film characters. 2014. 2

[5] Uttaran Bhattacharya, Gang Wu, Stefano Petrangeli, Viswanathan Swaminathan, and Dinesh Manocha. Highlightme: Detecting highlights from human-centric videos. In *ICCV*, 2021. 7

[6] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. 2017. 6

[7] Shixing Chen, Xiaohan Nie, David Fan, Dongqing Zhang, Vimal Bhat, and Raffay Hamid. Shot contrastive self-supervised learning for scene boundary detection. 2021. 2

[8] Hao Cheng, Zhaowei Zhu, Xingyu Li, Yifei Gong, Xing Sun, and Yang Liu. Learning with instance-dependent label noise: A sample sieve approach. In *ICLR*, 2021. 3

[9] Yashar Deldjoo, Mihai Gabriel Constantin, and Bogdan Ionescu. Mmtf-14k: A multifaceted movie trailer feature dataset for recommendation and retrieval. In *ACMMM*, 2018. 3

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Xiaohua Zhai Dirk Weissenborn, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. 2021. 6

[11] 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. 2021. 4

[12] Curtis G, Northcutt, Lu Jiang, and Isaac L. Chuang. Confident learning: Estimating uncertainty in dataset labels. 2021. 5

[13] Jort F. Gemmeke, Daniel P. W. 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. 2017. 6

[14] Michael Gygli, Yale Song, and Liangliang Cao. Video2gif: Automatic generation of animated gifs from video. In *CVPR*, 2016. 2, 7

[15] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *NeurIPS*, 2018. 3

[16] Fa-Ting Hong, Xuanteng Huang, Wei-Hong Li, and WeiShi Zheng. Mini-net: Multiple instance ranking network for video highlight detection. In *ECCV*, 2020. 2, 7

[17] Jinchi Huang, Lie Qu, Rongfei Jia, and Binqiang Zhao. O2u-net: A simple noisy label detection approach for deep neural network. In *ICCV*, 2019. 3

[18] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie understanding. In *ECCV*, 2020. 2, 3

[19] Qingqiu Huang, Yuanjun Xiong, Yu Xiong, Yuqi Zhang, and Dahua Lin. From trailers to storylines: An efficient way to learn from movies. *arXiv preprint arXiv:1806.05341*, 2018. 3

[20] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *ICML*, 2018. 3

[21] Yifan Jiao, Zhetao Li, Shucheng Huang, Xiaoshan Yang, Bin Liu, and Tianzhu Zhang. Three-dimensional attention-based deep ranking model for video highlight detection. In *IEEE Transactions on Multimedia*, page 2693–2705, 2018. 2

[22] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, page 1–1, 2019. 3

[23] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 6

[24] 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. 4, 6

[25] Jie Lei, Tamara L. Berg, and Mohit Bansal. Qvhighlights: Detecting moments and highlights in videos via natural language queries. In *NeurIPS*, 2021. 2

[26] Junnan Li, Richard Socher, and Steven C.H. Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. 2020. 3

[27] Shuaicheng Li, Feng Zhang, Kunlin Yang, Lingbo Liu, Shinan Liu, Jun Hou, and Shuai Yi. Probing visual-audio representation for video highlight detection via hard-pairs guided contrastive learning. *arXiv preprint arXiv:2206.10157*, 2022. 7

[28] Yang Liu and Hongyi Guo. Peer loss functions: Learning from noisy labels without knowing noise rates. 2020. 3

[29] 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. 2, 6, 7, 8

[30] Xingjun Ma, Hanxun Huang, Yisen Wang, Simone Romano, Sarah Erfani, and James Bailey. Normalized loss functions for deep learning with noisy labels. 2020. 3

[31] Curtis G. Northcutt, Lu Jiang, and Isaac L. Chuang. Confident learning: Estimating uncertainty in dataset labels. In *JAIR*, 2019. 3

[32] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, page 1345–1359, 2010. 2

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Sandhini Agarwal Gabriel Goh, Girish Sastry, Amanda Askell, Pamela Mishkin, and et al. Jack Clark. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 6

[34] 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. 2020. 2

[35] Yale Song, Yahoo Labs, New York, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. Tvsum: Summarizing web videos using titles. In *CVPR*, 2015. 2

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

[37] Min Sun, Ali Farhadi, and Steve Seitz. Ranking domain-specific highlights by analyzing edited videos. In *ECCV*, 2014. 2, 6, 7

[38] Zeren Sun, Fumin Shen, Dan Huang, Qiong Wang, Xiangbo Shu, Yazhou Yao, and Jinhui Tang. Pnp: Robust learning from noisy labels by probabilistic noise prediction. In *CVPR*, 2022. 3

[39] Zeren Sun, Yazhou Yao, Xiu-Shen Wei, Yongshun Zhang, Fumin Shen, Jianxin Wu, Jian Zhang, and Heng-Tao Shen. Webly supervised fine-grained recognition- benchmark datasets and an approach. In *ICCV*, 2021. 3

[40] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. 2016. 7

[41] Hao Tang, Vivek Kwatra, Mehmet Emre Sargin, and Ullas Gargi. Detecting highlights in sports videos: Cricket as a test case. In *ICME*, 2011. 2

[42] Paul Vicol, Makarand Tapaswi, Lluis Castrejon, and Sanja Fidler. Towards understanding human-centric situations from videos. 2018. 2

[43] Lezi Wang, Dong Liu, Rohit Puri, and Dimitris N. Metaxas. Learning trailer moments in full-length movies with co-contrastive attention. In *ECCV*, 2020. 2, 3, 7

[44] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. 2019. 3, 7

[45] Fanyue Wei, Biao Wang, Tiezheng Ge, Yuning Jiang, Wen Li, and Lixin Duan. Learning pixel-level distinctions for video highlight detection. In *CVPR*, 2022. 2, 7

[46] Hongxin Wei, Lei Feng, Xiangyu Chen, and Bo An. Combating noisy labels by agreement: A joint training method with co-regularization. In *CVPR*, 2020. 3

[47] 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 *CVPR*, 2022. 2, 3

[48] Xiaobo Xia, Tongliang Liu, Bo Han, Nannan Wang, Mingming Gong, Haifeng Liu, Gang Niu, Dacheng Tao, and Masashi Sugiyama. Part-dependent label noise: Towards instance-dependent label noise. In *NeurIPS*, 2020. 3

[49] Xiaobo Xia, Tongliang Liu, Nannan Wang, Bo Han, Chen Gong, Gang Niu, and Masashi Sugiyama. Are anchor points really indispensable in label-noise learning? In *NeurIPS*, 2019. 3

[50] Minghao Xu, Hang Wang, Bingbing Ni, Riheng Zhu, Zhenbang Sun, and Changhu Wang. Cross-category video highlight detection via set-based learning. In *ICCV*, 2021. 2, 6, 7

[51] Fengxiang Yang, Ke Li, Zhun Zhong, Zhiming Luo, Xing Sun, Hao Cheng, Xiaowei Guo, Feiyue Huang, Rongrong Ji, and Shaozi Li. Asymmetric co-teaching for unsupervised cross domain person re-identification. In *AAAI*, 2020. 3

[52] Shuo Yang, Erkun Yang, Bo Han, Yang Liu, Min Xu, Gang Niu, and Tongliang Liu. Estimating instance-dependent bayes-label transition matrix using a deep neural network. In *CVPR*, 2022. 3

[53] Ting Yao, Tao Mei, and Yong Rui. Highlight detection with pairwise deep ranking for first-person video summarization. In *CVPR*, 2016. 2

[54] Chuanyi Zhang, Yazhou Yao*, Xing Xu, Jie Shao, Jingkuan Song, Zechao Li, and Zhenmin Tang. Extracting useful knowledge from noisy web images via data purification for fine-grained recognition. In *ACMMM*, 2021. 5

[55] Zhilu Zhang and Mert R. Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *NeurIPS*, 2018. 3

[56] Guoqing Zheng, Ahmed Hassan Awadallah, and Susan Dumais. Meta label correction for noisy label learning. In *AAAI*, 2021. 3

[57] A. Zlatintsi, P. Koutras, N. Efthymiou, P. Maragos, A. Potamianos, and K. Pastra. Quality evaluation of computational models for movie summarization. 2015. 2