Temporal Sentiment Localization: Listen and Look in Untrimmed Videos

Zhicheng Zhang
gloryzzc6@sina.com
Nankai University
Tianjin, China

Jufeng Yang
yangjufeng@nankai.edu.cn
Nankai University
Tianjin, China

Figure 1: Sentimental segments from *The Wolf of Wall Street*. In the purple boxes, we show the classic segments (highlights of the film) that underpin the development of the story in the video, where each segment may convey a different sentiment. In the green boxes, the intervals between the highlights provide contextual information and carry neutral sentiment.

ABSTRACT

Video sentiment analysis aims to uncover the underlying attitudes of viewers, which has a wide range of applications in real world. Existing works simply classify a video into a single sentimental category, ignoring the fact that sentiment in untrimmed videos may appear in multiple segments with varying lengths and unknown locations. To address this, we propose a challenging task, *i.e.*, **T**emporal **S**entiment **L**ocalization (TSL), to find which parts of the video convey sentiment. To systematically investigate fully- and weakly-supervised settings for TSL, we first build a benchmark dataset named TSL-300, which is consisting of 300 videos with a total length of 1,291 minutes. Each video is labeled in two ways, one of which is frame-by-frame annotation for the fully-supervised setting, and the other is single-frame annotation, *i.e.*, only a single frame with strong sentiment is labeled per segment for the weakly-supervised setting. Due to the high cost of labeling a densely annotated dataset, we propose TSL-Net in this work, employing single-frame supervision to localize sentiment in videos. In detail, we generate the pseudo labels for unlabeled frames using a greedy search strategy, and fuse the affective features of both visual and audio modalities to predict the temporal sentiment distribution. Here, a reverse mapping strategy is designed for feature fusion, and a contrastive loss is utilized to maintain the consistency between the original feature and the reverse prediction. Extensive experiments show the superiority of our method against the state-of-the-art approaches.

MM '22, October 10–14, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3548007

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; • **Information systems** → **Sentiment analysis**.

KEYWORDS

dataset, video sentiment analysis, weakly-supervised learning

ACM Reference Format:
Zhicheng Zhang and Jufeng Yang. 2022. Temporal Sentiment Localization: Listen and Look in Untrimmed Videos. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22).* ACM, Lisboa, Portugal, 10 pages. https://doi.org/10.1145/3503161.3548007

1 INTRODUCTION

Video has become the most popular information carrier on Internet, *e.g.*, within the five minutes you've been reading the abstract, YouTube has received 2,500 hours of videos uploaded from users [49]. Therefore, it is naturally urgent to analyze such a large amount of video data. For this purpose, video content analysis focusing on objects and actions has been studied for decades. However, there still exists an affective gap between the content present in videos and the cognitive emotion of ones who view the videos [67]. Since video sentiment analysis aims to automatically reveal people's underlying attitudes toward a video [48], it meets this need and has a large number of applications in three aspects. First, content providers like YouTube could protect the mental health of teenagers from videos containing violent [7] or abnormal content [47, 53] via sentiment-based detection. Second, companies could utilize video data to develop product plans guided by commercial advertising analysis [17] and consumer sentiment analysis [9]. Last, online education platforms could improve the quality of teaching by automatically analyzing user's feedback and his/her attitude [1, 45]. Due

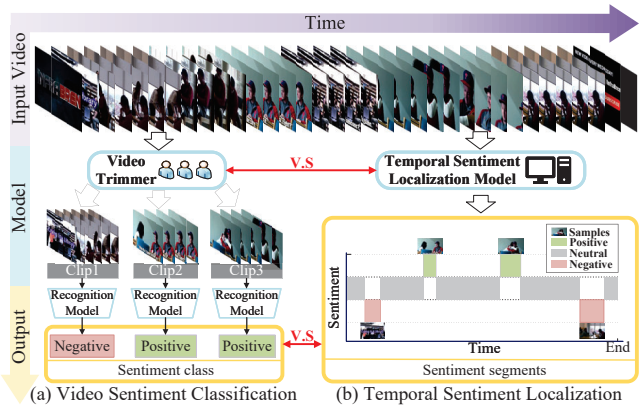

Figure 2: Difference between TSL and the previous works. (a) In video sentiment classification, the video is cropped manually, and multiple segments are then fed into the classification model. (b) TSL aims to simultaneously localize and classify multiple segments conveying sentiments in a video.

to the existence of various downstream applications, many efforts have been made in the field of video sentiment analysis [10, 31, 52].

Currently, most works focus on recognizing the single dominant sentiment evoked in a video [25, 60, 63]. They first crop the video into segments based on the speaker's sentence to construct the dataset [25, 60, 63]. Next, classification networks are developed to predict the sentiment of each segment [34, 37]. Thus, existing works that recognize the sentiment of trimmed video (*i.e.* cropped segments) have two limitations. First, cropping the video into segments is time-consuming and labor-sensitive, with human experts operating or requiring extra information such as subtitles [60, 63]. Second, recognizing the dominant sentiment in the trimmed video limits the downstream applications that require analyzing untrimmed videos such as movies and TV shows [7, 44]. Therefore, to avoid additional processes and support wider applications, some algorithms are developed to recognize the sentiment in untrimmed videos [19, 57, 71]. This is a challenging task because the sentiment conveyed in the untrimmed video become more complex with increasing length. For instance, as shown in Figure 1, a video may convey multiple sentiments and each sentiment appears with varying lengths and locations.

To address this, we propose **T**emporal **S**entiment **L**ocalization (TSL) to localize multiple segments conveying sentiments. Different from previous works that only predict sentimental category (Figure 2(a)), this work aims to simultaneously localize and classify the sentiments in untrimmed videos (Figure 2(b)). Specifically, with the prediction of all frames, we could aggregate the sentiment segments temporally and provide segment-level classification results. As is well known, training an effective model highly depends on a large-scale precisely labeled video dataset, *i.e.*, per-frame annotation. This labeling strategy is expensive due to the large number of frames. Moreover, since sentiment has the inherent characteristics of subjectivity and ambiguity, it usually needs multi-round labeling for this task. To alleviate the annotation cost, some weakly-supervised methods have been proposed [13, 21, 24]. The employed supervisions can be grouped into video-level and single-frame signals. Video-level signal [33, 36] indicates whether a class exists in the video, which is the weakest cue for training models. While single-frame signal [35, 38] additionally provides the timestamp of the occurrence for better localization performance. Besides, compared to full supervision that all the frames in the video are labeled, a single-frame supervision only provides one labeled frame per segment. Although employing weak supervision helps to alleviate heavy annotation burden (see details in Table 1), it brings new challenges. In this paper, we systematically investigate the fully- and weakly-supervised settings for TSL to reduce annotation cost.

First, we build a benchmark dataset, namely TSL-300. Specifically, we collect 300 untrimmed videos with a total length of 1,291 minutes from the well-used datasets in this field, *i.e.*, Ekman6 [57], VideoEmotion8 [19], CMU-MOSI [62], CMU-MOSEI [62]. The videos include movies, TV shows, and speech videos. Each video is additionally labeled in a frame-by-frame manner for fully-supervised learning setting, as well as single-frame annotation for weakly-supervised learning setting. Second, we propose a weakly-supervised multimodal fusion network, called TSL-Net, which fuses multimodal features from both visual and audio modalities. To preserve the information from each affective modality, we develop a reverse mapping algorithm to predict the feature before fusion. We then employ a contrastive loss to regularize the consistency between the original features and the reverse prediction. Using two training signals generated from single-frame annotations, we optimize the network by jointly training. For the frame-by-frame signal, we develop it to represent the sentiment within each frame. Based on confidence of frame-by-frame predictions, we search for pseudo labels to complete the signal via a greedy search strategy. For the video-level signal, we use it to represent the conveyed sentiment of a video, where we model the sentiment distribution guided by the distribution of single-frame annotations.

The contributions of this work are three-fold: 1) We propose a challenging task of temporal sentiment localization to bridge the gap between video sentiment analysis and real-world applications. A benchmark dataset named TSL-300 is collected, where both the frame-by-frame annotations and single-frame annotations are provided. The dataset will be released to the community and may boost the research in this field. 2) We present a weakly-supervised framework, *i.e.*, TSL-Net, which can train a multimodal fusion network for temporal sentiment localization only using the single-frame supervision. This setting reduces the cost of dense annotations. 3) Extensive experiments demonstrate the superiority of our method against the state-of-the-art weakly-supervised methods.

2 RELATED WORK

2.1 Video Sentiment Analysis

Dataset. Due to the diversity of types of videos, video sentiment analysis has investigated multiple carriers of the sentiment, such as speech videos and movies. According to the type of video data, existing datasets for video sentiment analysis can be grouped into trimmed video datasets [62, 63] and untrimmed video datasets [19, 25, 57]. To analyze the sentiment in the trimmed video, CMU-MOSI [62] and CMU-MOSEI [63] datasets collect 2,199 and 23,453 sentences in the video, respectively. The sentences are cropped

from the video based on the punctuation of video transcription and then verified by human experts. In the CH-SIMS dataset [60], human experts use video editing tools manually for precise cropping. To directly predict the sentiment of untrimmed videos from social media, VideoEmotion8 [19] and Ekman6 [57] datasets collect 1,101 and 1,637 web videos from YouTube and Flickr, respectively. With an average duration of 107 and 112 seconds, two datasets are labeled for single-label classification.

Method. Video sentiment analysis methods [34, 37, 71] leverage multiple affective cues such as visual, audio, and textual information [56] for predicting sentiment, which can be split into direct methods and indirect methods. Early works [16, 20, 22, 54] directly estimate the conveyed sentiment of the videos. For example, Kang [22] maps the low-level features to the sentiment of the movies by hidden markov models. To improve the effectiveness of the extracted features, Jiang *et al.* [19] leverage the rich semantic information of adjective-noun pairs to build up mid-level representation from SentiBank [2]. Due to the high-level abstract of sentiment, some indirect methods [25, 68, 68] focus on leveraging the emotional impact of the video contents, such as the facial expression of speakers. The key to these methods is the fusion of multimodal information, including facial expression, audio sentiment, and so on. To capture the information in each affective modality, CapsGCN [31] integrates inter-relations and intra-relations of multimodality by graph convolutional network.

2.2 Video Temporal Localization

Fully-supervised temporal localization [12, 43, 50, 64] is to recognize the class of instances and locate the corresponding temporal boundaries in the untrimmed video, which can be divided into two patterns, *i.e.*, one-stage [32] and two-stage [5, 8, 58]. The difference is that two-stage methods generate proposals and then classify the categories. Although fully-supervised temporal localization methods achieve promising performance, the model highly relies on densely annotated data, *i.e.* labeling thousands of frames in a video (see details in Table 1).

Weakly-supervised temporal localization [11, 14, 18, 26, 41, 46, 66, 73] aims to weaken the requirement for annotations, *i.e.* train a model by the easily accessible supervision signal. According to the used weak supervision signals, existing methods can be roughly grouped into video-level and single-frame. For the methods utilizing video-level supervision, UntrimmedNet [55] is the first to learn the temporal action localization network by coupling the classification and selection modules. However, there exists ambiguity mapping problem when multiple actions are presented in a video [72]. To address this, CoLA [66] introduces a mining strategy to locate the snippets and integrates a contrastive loss at the snippet level. More recently, to better leverage the annotation, single-frame supervision [38] additionally provides a timestamp per segment to indicate the presented action. These methods generate the pseudo label guided by the timestamp information to train the model. According to the contextual frame and pseudo background frame, SF-Net [35] identifies the pseudo action in unlabeled frames. Further, LACP [27] leverages the background points to supplement the action, and then learns representations via feature similarity.

3 TSL-300 DATASET

Our dataset consists of 300 untrimmed videos with an average video time of 4.3 minutes. We filter out three types of abnormal videos to provide a high-quality and diverse dataset. The selected videos are annotated by four well-trained annotators from different backgrounds. In order to alleviate the burden of dense annotation, we study the weakly-supervised setting and the fully-supervised setting used for training a model.

3.1 Data Collecting

To build up our dataset, the types of collected videos include speech videos, movies, and TV shows. Specifically, we collect speech videos from the raw untrimmed data of the CMU-MOSEI [63] and CMU-MOSI [62]. The movies and TV shows are derived from the Ekman6 [57] and VideoEmotion8 [19] datasets. We first collect 6,667 untrimmed videos. Then, we filter out two types of abnormal data during the checking and selecting stages. For one thing, sequences of static images are not taken into consideration since they have been well studied in visual sentiment analysis. For another thing, the recurring videos reduce diversity by conveying similar sentiments over a repeated duration.

3.2 Data Annotation

Due to the subjectivity of sentiment, different people might feel differently about the same video. To avoid bias from the annotators, we select four annotators from different backgrounds.

Before annotating, we train and test the annotators to guarantee the quality of the labels. In order to train annotators, we provide annotation guidelines to all participants, including the purpose of our study, detailed descriptions of the annotation process, the statistics of our dataset, and instructions of labeling software. For testing, we ask each participant to annotate 10 videos covering all types of videos. We regard a label as correct when the category is labeled correctly and the segment possesses more than half of the overlap with the label. An annotator will not be hired before he or she achieves an accuracy of 80% in the testing.

During annotation, we use the sentiment model from philosophical theory [6], which groups human attitudes toward instances into positive, neutral, and negative attitudes. Then, we aggregate the annotations from multiple annotation professors. With the aggregated label, the proportion of achieving agreement (*i.e.* over half annotators vote for the same sentiment) is 98.97% for TSL-300.

Frame-by-frame annotation. For providing dense annotations, *i.e.*, full supervision, an annotator follows the below scheme to annotate a video multi-round. First, the annotator labels a segment when it evokes an emotional response. Then, aiming to indicate a segment, the temporal boundary and its corresponding sentiment are recorded. Finally, the multi-round labeling ends when an annotator does not find a new sentiment after watching the video. Considering the subjectiveness of annotators, we aggregate the annotations from multiple annotators. We yield the label of each frame according to each annotator and then aggregate sentiment label by majority voting for frame-by-frame annotation.

Single-frame annotation. To investigate how to alleviate the annotation burden, we label these videos in a weakly-supervised setting. The aforementioned full supervision requires annotators to

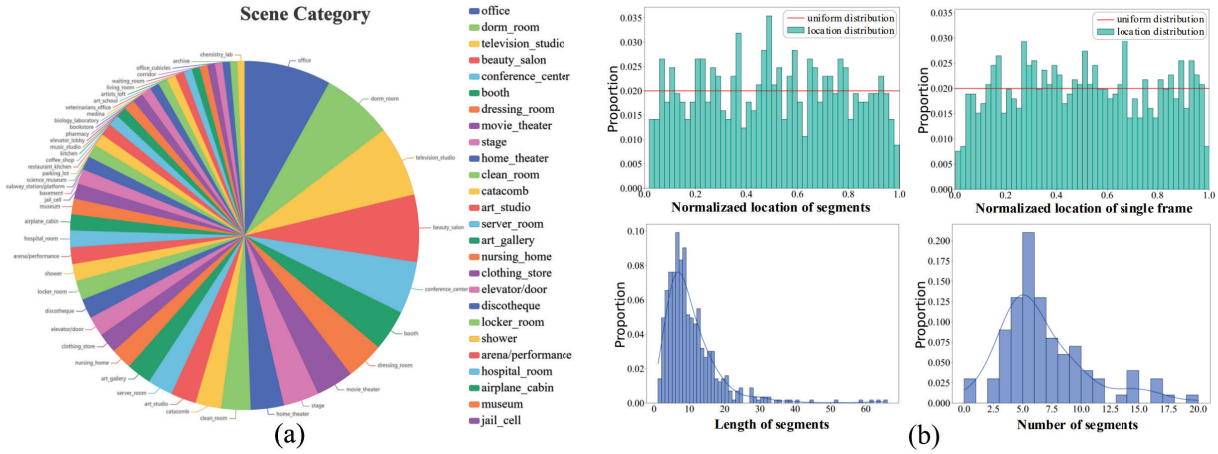

Figure 3: The statistics around collecting and labeling of our dataset. (a): The distribution of scene category is given by PlacesCNN [74]. (b): The distribution of segments and the single-frame annotaions in our TSL-300 dataset, including number, location and length.

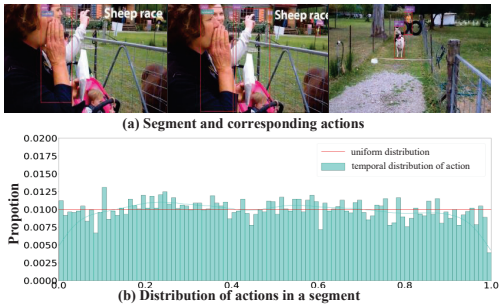

Figure 4: The statistics around actions of our dataset. (a): A segment of *Sheep Racing* and corresponding actions. (b): The temporal location distribution of action in our TSL-300 dataset.

watch a video multiple times and carefully determine the temporal boundary of each segment. To reduce the time cost, we annotate the sentiment label only within the video's keyframes, which directly evoke the viewer's strong sentiment response [71]. Specifically, the annotators pause the video when they are aroused and annotate the paused frame with the corresponding sentiment. Again, we aggregate the annotations from multiple annotators as labels [35].

Compared to the full supervision, the single-frame annotation sharply decreases the annotation cost, since the number of frames need to be labeled is hundreds of times smaller than the frame-by-frame annotation (Table 1). Besides, similar to the most commonly used video-level supervision [65, 66], the single-frame supervision also requires viewing the whole video and then assigning the labels for the video. While such supervision can additionally provide the temporal location information (*i.e.* the timestamp of labeled frame) with few extra annotation burden.

3.3 Statistics of TSL-300

In our dataset, the untrimmed videos are collected from TV shows, movies, and speech videos. After filtering out abnormal data, we collect 300 untrimmed videos, including half of the speech videos and half of the TV shows and movies. The average and maximum length of the collected video is 258.2 and 1,160.1 seconds, respectively. These videos are split into 200 videos for training and 100 videos for testing. As a result, our dataset contains of 1,642 sentiment segments for a total of 1,291 minutes. Here, we demonstrate the statistics of our dataset from the aspect of data collecting, annotation, and characteristic.

The videos of TSL-300 are collected from multiple sources, such as pushing a YouTuber live stream in the office and reporting news in the television studio. The involved scenes of collected videos vary from the office to the museum and the types of scenes range from indoor to outdoor. We leverage the power of the scene understanding method [74] to demonstrate the statistics of the collected scene. As demonstrated in Figure 3(a), the top 3 most appeared scenes are office, dorm room, and television studio.

We plot the statistic about annotations in Figure 3(b). The left top figure depicts the segment location in the video. As can be seen, the distribution of location is close to the uniform distribution, while achieving the maximum in the middle. Because the middle part of the video often represents the climax of the story, which tends to convey rich sentiment. The right top figure shows the single-frame annotations' location in the segment. Similarly, the distribution is close to the uniform distribution. The bottom two figures show the distribution of sentiment segments in TSL-300. Compared to other video datasets, the number of segments in our TSL-300 (5.4) is larger than ActivityNet [3] (1.5) and HACS [69] (2.8), while smaller than THUMOS14 [15] (15.4). The reason is that there is a gap between recognition and perception, *i.e.*, each conveyed sentiment is caused by multiple actions [67].

For the characteristic of video, we visualize the sentiment segment and plot the distribution of actions in Figure 4, generated by the action detector [51]. First, as can be seen, a sentiment segment usually contains multiple actions. Therefore, a sentiment segment could transmits rich sentiment to the viewers. Besides, this is consistent with the fact that the length of sentiment segment is longer than action. Second, we notice that the actions appear evenly in the sentiment segment. Considering sentiments are correlated with actions, *e.g.*, the situations of characters [52], it shows the labeled segments have consistency in selecting the temporal boundary.

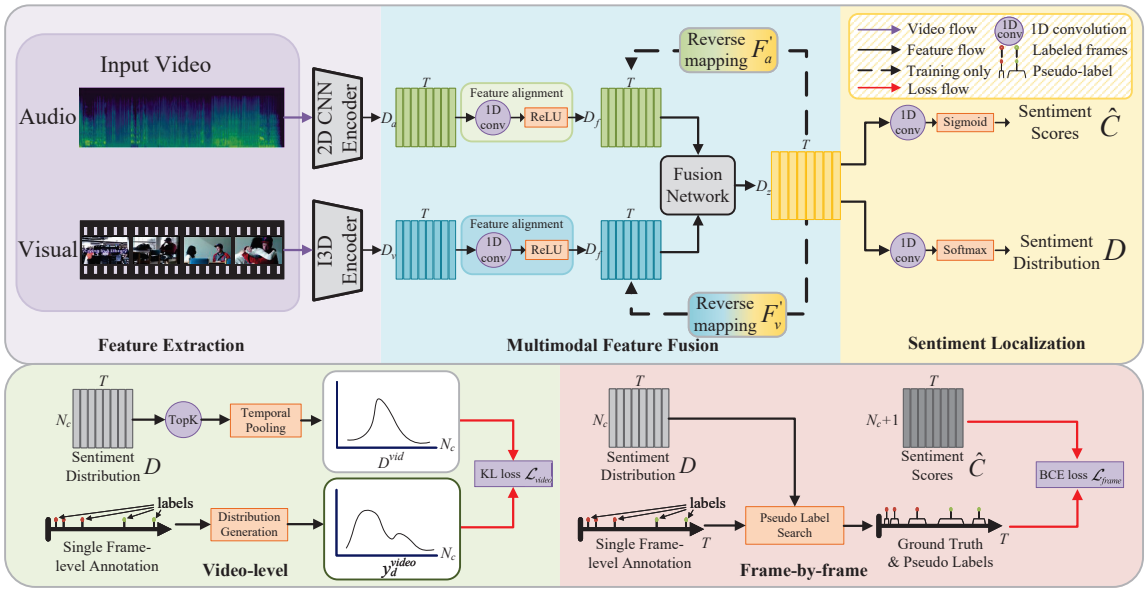

Figure 5: The pipeline of our method. The input is an untrimmed video consisting of visual modality and audio modality. Our multimodal fusion network fuses the features from two modalities to predict the sentiment of each frame for localization. The reverse mapping function is employed to maintain the similar semantic for feature fusion during training. The network is optimized by jointly training from the generated video-level and the searched frame-by-frame signals.

4 METHODOLOGY

4.1 Formulation

For each training video v with T frames, the corresponding labels in labeled frames are $\boldsymbol{y} = \{(t_1, cls_1), \ldots, (t_{N_k}, cls_{N_k})\}$, where t_i is the timestamp that indicates the i^{th} labeled frame and $cls_i \in \mathbb{R}^{N_c}$ denotes the one-hot vector, N_k and N_c represent the number of labeled frames and sentiment class, respectively. The labels are ordered according to the time that $t_{i-1} < t_i < t_{i+1}$. For a testing video, we aim to predict segments conveying sentiment $\{(t_s, t_e, cls, \phi)\}$, where t_s, t_e, cls, ϕ denote start, end, class, and confidence of the segment, respectively. Here, the labels of testing videos are the set of sentiment segments $\{(t_s, t_e, cls)\}$.

4.2 Network Architecture

To address TSL, we design a weakly-supervised framework to leverage the nature of the video, *i.e.*, multimodal information and ambiguity of sentiment labels, as demonstrated in Figure 5.

Feature Extraction. Since video naturally contains rich affective cues from multiple modalities, we extract features of visual and audio modalities following [71]. For the visual modality, we split the videos into 16-frame RGB segments and feed them into the pre-trained I3D extractor [4], resulting in the visual feature $X_v \in \mathbb{R}^{D_v \times T}$. Here, T and D_v represent the length and feature dimensions, respectively. For the audio modality, we use the most well-known descriptor Mel-Frequency cepstral coefficients (MFCC). Besides, the audio description is further processed by a 2D CNN to yield audio features $X_a \in \mathbb{R}^{D_a \times T}$. D_a is the feature dimension.

Multimodal Feature Fusion module. Both visual and audio modalities are included for predicting the sentiment in each timestamp [70]. However, a direct fusion of these features may lead to the overfitting of the model. Therefore, we propose a fusion network to build multimodal representation f_z. To fill up the semantic gap between two modalities, the features X_a and X_v are aligned by a 1D convolution with ReLU activation. Thus, we yield $f_m \in \mathbb{R}^{D_f \times T}$ for fusion, where D_f denotes the dimension of aligned feature and $m \in \{v, a\}$ denotes the index of modality. Sequentially, the two features are fused by a subnet $F_m : \mathbb{R}^{D_f} \rightarrow \mathbb{R}^{D_z}$ along the time dimension. For solving numerical instability, inspired by CPC [42], we ensure that the fused features maintain their previous patterns. Specifically, nonlinear mappings $F'_m : \mathbb{R}^{D_z} \rightarrow \mathbb{R}^{D_f}$ are applied to predict the original feature from each modality during training. Besides, to avoid the trivial solution of directly preserving the original one, *i.e.* $F'_m(f_z) = f_m$, we take the cosine similarity as the distance $D(x, y) = \frac{xy^T}{||x|| \cdot ||y||}$, where $||\cdot||$ is the normalization operation. Then, we compute the loss as

$$\mathcal{L}_{fuse} = -\sum_{m \in \{v,a\}} \sum_{t=1}^{T} \log \frac{D\left(f_m^t, F'_m(f_z^t)\right)}{\sum_{i \neq t} D\left(f_m^t, F'_m(f_z^i)\right)}. \quad (1)$$

Sentiment Localization. With the fused features, we aim to classify foreground/background and different sentiments in the frames. The classifier consists of a 1D convolution and a sigmoid function, where we aim to predict the class of foreground/background and each sentiment class $Logits \in \mathbb{R}^{T \times N_c}$. N_c is the number of sentiment classes. Here, the probability of foreground/background indicates whether the frame conveys sentiment or not. Meanwhile, the classifier gives the scores of different sentiment classes. For clarity, we denote the probabilities of foreground/background and multiple sentiment classes as $A \in \mathbb{R}^T$ and $C \in \mathbb{R}^{T \times N_c}$.

Based on the prior of foreground, the probability $\widehat{C}$ of the c^{th} sentiment class in the t^{th} frame can be calculated as $\widehat{C}_c^t = C_c^t \cdot A^t$. To

learn the intensity in different sentiments, we derive the corresponding distribution $D \in \mathbb{R}^{T \times N_c}$ before the sigmoid function. Further, we can generate the video-level prediction C_c^{vid}, D_c^{vid} based on the frame-by-frame predictions. Following [29], we adopt temporal k pooling [40] as

$$C_c^{vid} = \frac{1}{k} \sum_{t \in U} \widehat{C}_c^t, \quad U = \arg topk(\widehat{C}_c), \tag{2}$$

where k is set as $T/8$ for the top-k operation. The predicted probabilities are then used to compute the losses, *i.e.*, $\mathcal{L}_{frame}$ and $\mathcal{L}_{video}$.

4.3 Learning from Single-frame Supervision

Video-level. With single-frame annotations, we can generate video-level labels for effective training [35]. To generate the labels in an untrimmed video, we consider two types of information we need in training a model to locate the sentiment. Indeed, the probability could indicate whether a sentiment class exists, and the distribution can help to determine the intensity of each class. Specifically, we generate $y_d^{video} = \{d_1, \ldots, d_{N_c} | d_c \in [0, 1]\}$ in three ways: 1) Hard: we directly use the appearance of class to model the magnitude. The supervision signals are formulated into multi-hot labels as

$$d_c = \frac{\mathbf{1}(I_l^c = 1)}{\sum_c \mathbf{1}(I_l^c = 1)}, \tag{3}$$

where I_l^c is the class of the labeled frame and $\mathbf{1}(\cdot)$ is the indicator function, which is set as 1 if satisfying the conditions described above. 2) Label Smooth [39]: To help the model learn the magnitude of the multiple classes that exist simultaneously, we adopt label smoothing as

$$d_c = \begin{cases} 0.9, & c = \arg\max \sum_U \mathbf{1}(I_l^c = 1) \\ 0.1, & c \neq \arg\max \sum_U \mathbf{1}(I_l^c = 1) \end{cases}, \tag{4}$$

where U is the set of the labeled frames. 3) Label-based: Considering the number of annotations in each class, we model the magnitude of the distribution according to the number of labeled frames I_l belonging to the c^{th} class,

$$d_c = \frac{1}{|U|} \sum_U \mathbf{1}(I_l^c = 1). \tag{5}$$

To address the ambiguity of intensity in sentiment classes, we employ the Kullback-Leibler divergence loss as

$$\mathcal{L}_{video} = -\sum_{c=1}^{N_c} d_c \ln D_c^{vid}. \tag{6}$$

Frame-by-frame. In the labeled training set $X_l = (t_i, cls_i)$ with N_l frames, the i^{th} labeled frame at timestamp t_i is assigned to a sentiment class cls_i. This allows us to directly compute losses when training the foreground classifier and the class classifier. Specifically, Binary Cross-Entropy loss in the form of Focal loss [30] is adopted to optimize the model in the labeled (foreground) frames. Formally, for the predicted class probability $\widehat{C}_c^t$ and foreground probability A^t, the loss is computed as

$$\mathcal{L}_{frame}^l = -\sum_{c=1}^{N_c} \left(cls_i \left(1 - \widehat{C}_c^{t_i}\right)^\beta \log \widehat{C}_c^{t_i} + (1 - cls_i)\, \widehat{C}_c^{t_i \beta} \log\left(1 - \widehat{C}_c^{t_i}\right) \right) + A_t^\beta \log 1 - A_t. \tag{7}$$

Due to only a single frame being annotated in a sentiment segment, the labeled frames are fewer than the total frames of the video (around a hundred times less in a video). To address this problem, existing methods search for available positive sample frames to supplement training. Inspired by [29], we propose a confidence-based search strategy to mine the sentiment in the foreground and background frames. Practically, according to the annotation scheme, there are background frames that exist between two labeled frames. The search strategy regards frames as background when they have low confidence between two labeled frames. However, learning from pseudo labels for background frames is not enough to train an effective model. To this end, we generate pseudo labels for unlabeled frames, including foreground and background. Specifically, we assign foreground frames based on confidence during training. We select the frames with high confidence between the two labeled frames and then assign their labels according to the nearest label.

The generated pseudo-labels lead to the problem that the labels are wrong in the early stages of model training, especially when there exists affective gap [70] between high-level sentiment and deep representations. Further, learning from these labels may lead to the overfitting problem. To this end, we propose to learn from the soft pseudo label, which is easier to generate. Specifically, for the unlabeled training set $X_{ul} = \left(t_i, y_i^{ul}\right)$ with N_{ul} frames, the pseudo distribution of i^{th} frames are generated as $y_i^{ul} = \{d_c^{ul} \in [0, 1], c \in \mathbb{R}^{N_c+1}$. In the foreground frames, we generate distribution as pseudo label from the prediction of labeled frames. While for the background one, the distribution is the prediction of the background segment with maximum confidence. Naturally, we could compute the corresponding loss in the foreground and background frames as

$$\mathcal{L}_{frame}^{ul} = -\sum_{c=1}^{N_c} \left(d_c^{ul} \left(1 - \widehat{C}_c^{t_i}\right)^\beta \log \widehat{C}_c^{t_i} + \left(1 - d_c^{ul}\right) \widehat{C}_c^{t_i \beta} \log\left(1 - \widehat{C}_c^{t_i}\right) \right) + d_{N_c+1}^{ul} (1 - A_t)^\beta \log A_t + (1 - d_{N_c+1}^{ul}) A_t^\beta \log (1 - A_t). \tag{8}$$

Then, we compute the frame-by-frame loss with coefficient μ as

$$\mathcal{L}_{frame} = \mu \mathcal{L}_{frame}^l + (1 - \mu) \mathcal{L}_{frame}^{ul}. \tag{9}$$

Overall, our training objects of multi-level joint training framework can be formulated with coefficients λ_1, λ_2 as

$$\mathcal{L} = \mathcal{L}_{fuse} + \lambda_1 \mathcal{L}_{video} + \lambda_2 \mathcal{L}_{frame}. \tag{10}$$

5 EXPERIMENT

5.1 Evaluation Metrics

For evaluation, we follow the commonly used metric mAP@tIoU in temporal localization [3, 27, 28, 35, 66]. Under different intersection over union (IoU), the metric is calculated by mean Average Precision (mAP). Due to the challenge of high-level abstract, we set thresholds ranging from 0.1 to 0.3, with an interval of 0.05. As TSL is in the early stages of downstream applications, we prefer recall over precision. Thus, we report Recall and F2-score as $F_\beta = \frac{(1+\beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall}$, where β is 2. We also report averages of these values.

5.2 Implementation Details

The visual frames are split into 16-frame clips via zero-padding, which are then fed into the corresponding feature extractor. Besides,

Table 1: Quantitative performance of the proposed method and SOTA methods in our dataset. We include the methods using frame-by-frame full supervision and video-level weak supervision for reference. † indicates that the method uses the same feature extractor as ours. Average denotes the mean values of metric at IoU threshold 0.1:0.05:0.3. Bold and underline indicate the highest and second-highest performance, respectively. We report the annotation cost of labeling 200 training videos for three types of annotation.

Supervision	Annotation Cost	Method	Year	mAP@IoU(%)					Average		
				0.1	0.15	0.2	0.25	0.3	mAP	Recall	F2 score
Full Supervision	80h:41m	VAANet [71]	2020	28.01	24.95	20.68	16.45	12.50	20.51	69.82	39.59
		AFSD [29]	2021	16.95	13.74	12.23	11.21	09.41	12.70	62.30	30.17
		AFSD† [29]	2021	30.12	26.74	23.83	21.25	20.51	24.49	73.58	41.27
Weak Supervision (Video-level)	00h:54m	CoLA [66]	2021	11.17	09.33	07.87	06.34	05.13	07.96	63.02	18.20
		UM [28]	2021	13.49	11.69	09.90	07.79	05.82	09.73	44.67	22.42
		CoLA† [66]	2021	11.87	10.14	08.11	06.83	04.98	08.38	64.07	21.19
		UM† [28]	2021	14.77	13.28	10.77	08.41	06.54	10.75	55.33	23.94
Weak Supervision (Single-frame)	01h:13m	LACP [27]	2021	12.82	11.04	09.32	07.26	06.06	09.30	60.31	22.91
		SFNet [35]	2020	12.73	10.47	08.12	06.81	04.38	08.50	42.17	19.93
		LACP† [27]	2021	17.25	14.46	12.47	09.59	07.76	12.30	66.92	25.01
		SFNet† [35]	2020	20.26	18.09	15.01	12.38	09.76	15.10	59.55	32.61
		Ours	2022	**28.72**	**24.92**	**20.46**	**16.10**	**11.83**	**20.40**	**71.14**	**35.36**

only the visual feature extractor is pretrained on Kinetics-400 [4]. To train the model, Adam [23] is adopted for optimization, with a learning rate of 10^{-5}. The hyperparameters are obtained by the grid search. The threshold for video-level prediction is set to 0.5. Following [27], we set the window size for temporal top-k pooling as one-eighth the length of the video. All of the experiments are performed on two 3090 GPUs with 24GB GPU memory.

5.3 Quantitative Analysis

In Table 1, we conduct experiments with three groups of methods on our dataset. Note that the fully-supervised methods utilize annotations of all frames, which is much more expensive than single-frame annotations. To show the required labor of annotation, we report the annotation cost for labeling these supervisions. The length of the videos is expelled since they are necessary for viewing.

Comparison with State-of-the-art Methods. We compare our TSL-Net with other weakly-supervised temporal localization methods. First, our method achieves competitive performance in comparison to the previous best method of SFNet, outperforming it for a margin of 33.43% in eight metrics. We give detailed reasons in our ablation studies. Second, for the case of low IoU thresholds, our method obtains a large performance improvement, *e.g.*, 41.7% in mAP@0.1 and 37.7% in mAP@0.2 against SFNet. Because frame-by-frame training searches for the labels of unlabeled frames could mine the potential sentiment. Third, the compared methods gain better performance when using the same audio feature extractor as ours. This confirms that leveraging multiple affective modalities can boost the performance of all methods.

Single-frame Supervision vs. Video-level Supervision. With 35.1% of increasing annotation cost, single-frame methods get better performance than video-level methods, in terms of eight metrics of 42.3%. We notice that single-frame methods perform better at a high IoU threshold, *i.e.* 11.83 of single-frame methods v.s 6.54 of video-level methods in mAP@0.3. This verifies that single-frame supervision can provide temporal location information of segments.

Single-frame Supervision vs. Full Supervision. As can be seen, our proposed single-frame methods achieve competitive performance with fully-supervised methods. For example, compared to VAANet, ours achieves competitive performance using only a few labeled frames. For clarity, VAANet selects keyframes of video and extracts corresponding features for classification. Therefore, it can localize the sentiment temporally by the selected keyframes. The reason for competitive performance is that we utilize a greedy search strategy to assign labels for the unlabeled frames and model the sentiment over the entire video to complement supervision signals. Thus, we bridge the gap between full supervision and single-frame one to some extent. Besides, AFSD gets slightly better performance due to more complete information of full supervision.

5.4 Ablation Studies

Modules & losses. To investigate video-level distribution learning, single frame sentiment learning, and multimodal feature alignment module, we conduct ablation studies as shown in Table 2. We further discuss each component by analyzing its variants.

How to generate video-level distribution? In Table 3, we aim to investigate the way of generating the video-level distribution. Here, we remove multimodal fusion and frame-by-frame training to build our baseline. First, we find that label-based distribution generation achieves the best performance. Because it leverages the magnitude relation between multiple sentiment classes. Intuitively, the magnitude of the sentiment conveyed is directly proportional to the time it occurs. Second, the Kullback-Leibler divergence loss achieves better performance at low IoU thresholds compared to the binary cross-entropy loss, such as 27.0% in mAP@0.1 and 45.7% in mAP@0.2. This is because the KL loss helps to mine the potential sentiment by simulating the corresponding magnitude impacted by the ambiguity of sentiment [59].

How to supervise the prediction of each frame? To investigate how to supervise the prediction frame-by-frame, we conduct ablation experiments as shown in Table 4. First, in comparison of lines 1 and 2, we notice a sharp decrease in performance when removing the loss for pseudo-label. This is because the number of labeled frames is small, and thus it is hard to optimize the model. Second, in comparison of lines 2 and 3, we find a large performance gap between using the hard CE loss in LACP and ours, especially at low IoU thresholds. Because our proposed loss for unlabeled frames reduces the ratio of the hard case to the learning objects, *i.e.* focuses on the case in the center of the segment, it alleviates the risk of

Table 2: Ablation studies on the modules of our method, where Vid, FbF, and MA denotes video-level training, frame-by-frame training, and multimodal feature fusion, respectively.

Module			mAP@IoU(%)			mAP
Vid	FbF	MA	0.1	0.2	0.3	AVG
✔			22.45	17.08	11.20	16.91
✔		✔	25.32	18.03	11.35	18.23
✔	✔	✔	**28.72**	**20.46**	**11.83**	**20.33**

Table 3: Ablation studies on the impact of video-level training.

Variant	mAP@IoU(%)			mAP
	0.1	0.2	0.3	AVG
BCE loss	16.37	11.72	07.34	11.81
Hard (Equ.3)	20.08	15.72	09.96	15.25
Smooth (Equ.4)	20.89	14.50	08.50	14.63
Label-based (Equ.5)	**22.45**	**17.08**	**11.20**	**16.91**

Table 4: Ablation studies on the impact of frame-by-frame training.

Variant	mAP@IoU(%)			mAP
	0.1	0.2	0.3	AVG
No pseudo-label	18.48	13.26	08.57	13.43
LACP [27]	16.37	11.72	07.34	11.81
Ours	**28.72**	**20.46**	**11.83**	**20.33**

Table 5: Ablation studies on the impact of multimodal feature fusion.

Variant	mAP@IoU(%)			mAP
	0.1	0.2	0.3	AVG
TFN [61]	18.60	12.78	08.72	13.36
No feature alignment	26.44	16.58	10.63	17.77
Use L2 dis in $\mathcal{L}_{fuse}$	25.56	18.22	10.90	18.22
Ours	**28.72**	**20.46**	**11.83**	**20.33**

Table 6: Ablation studies on the impact of affective modalities.

Metrics	Audio			Visual			Visual+Audio		
	[36]	[27]	Ours	[36]	[27]	Ours	[36]	[27]	Ours
mAP	07.41	08.18	**09.71**	08.50	09.30	**11.16**	15.10	12.30	**20.40**
Recall	36.37	54.70	**56.85**	42.17	60.31	**60.55**	59.55	66.92	**71.14**
F2 score	17.10	20.42	**27.37**	19.93	22.91	**26.33**	32.61	25.01	**35.36**

learning the wrong pseudo label. In contrast, LACP focuses on the hard samples in the margin of segments.

How to perform multimodal feature fusion for predicting sentiment? In Table 5, we conduct experiments to find the optimal design for multimodal feature fusion. First, to disable the alignment between visual modality and audio modality, both feature alignments $\mathbb{R}^{D_m} \rightarrow \mathbb{R}^{D_z}$ are removed. Then, we replace the cosine distance function with the inner product distance in $\mathcal{L}_{fuse}$, *i.e.* $D(x, y) = xy^T$. As can be seen, TFN which directly fuses multimodal features gets a suboptimal performance. This is consistent with our motivation for designing our fusion module. Among variants of our design, removing the feature alignment sharply decreases performance because it helps to learn the intrinsic part of unimodal features. The feature from multiple modalities can help localize the sentiment in the temporal dimension, which is verified by the observation that using our multimodal feature extractor can boost the performance, *i.e.*, achieving 27.7% improvement over SFNet, LACP, UM, CoLA in eight metrics.

Which affective modality contributes most? In Table 6, we analyze the effectiveness of each affective modality. We show the performance of our method and two SOTA methods using audio-only (A), visual-only (V), and visual-audio features (VA). We find

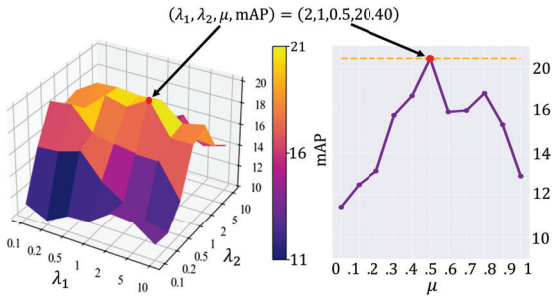

Figure 6: The hyperparameters analysis of λ_1, λ_2, and μ. We search for the best solution for average mAP.

that ours achieves competitive performance against other SOTA methods, in terms of 18.8% (A), 11.7% (V), and 37.8% (VA) improvement against the previous SOTA method. We also observe that visual modality contributes more than audio modality. It lies in that visual modality provides more high-level emotional cues, e.g. facial emotion and scene.

5.5 Hyperparameters Analysis

In Figure 6, we study the impact of λ_1, λ_2, and μ, in terms of the average of mAP@0.1:0.3. Thus, we conduct experiments to find the optimal values of λ_1 and λ_2 from $\{0.1, 0.2, 0.5, 1, 2, 5, 10\}$. For promising performance, we empirically set λ_1, λ_2 as 2, 1 to balance the influence between two losses. We find that too large and too small values of λ_1 and λ_2 can lead to degradation of the performance. To balance the effect between the losses for labeled part and unlabeled part, we search for the best value of μ from 0 to 1 with the interval of 0.1. Our method achieves optimal performance when treating the losses for labeled and unlabeled data equally. When considering only one of these two losses, *i.e.* $\mu = 0/1$, the model performs worse than using both. This is because the two complement each other. The unlabeled part increases the number of samples, and the labeled part learns reliable labels.

6 CONCLUSIONS

To narrow the gap between video sentiment analysis and real-world applications, we propose a new task, *i.e.* **T**emporal **S**entiment **L**ocalization, which will be useful for future studies in video sentiment analysis. A benchmark dataset containing 300 videos is built for fully- and weakly-supervised settings. To reduce the annotation cost, we propose a weakly-supervised method to utilize single-frame annotations for joint training. Our experiments raised two interesting observations: 1) Utilizing multimodal information can boost the performance of TSL. We speculate that visual and audio are highly related to sentiment and that affective modalities provide rich cues for sentiment localization. 2) Learning from sentiment distribution can help detect high-level and abstract sentiment. We believe learning from distribution can bridge the affective gap deriving from sentiment.

ACKNOWLEDGMENTS

This work was supported by the National Key Research and Development Program of China Grant (NO. 2018AAA0100403), NSFC (NO.61876094, U1933114), Natural Science Foundation of Tianjin, China (NO.20JCJQJC00020)

REFERENCES

[1] R Baragash and H Aldowah. 2021. Sentiment analysis in higher education: a systematic mapping review. In *Journal of Physics: Conference Series*.

[2] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. 2013. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *ACM MM*.

[3] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*.

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

[5] Yu-Wei Chao, Sudheendra Vijayanarasimhan, Bryan Seybold, David A Ross, Jia Deng, and Rahul Sukthankar. 2018. Rethinking the faster r-cnn architecture for temporal action localization. In *CVPR*.

[6] Julien Deonna and Fabrice Teroni. 2012. *The emotions: A philosophical introduction*. [Book].

[7] Florian Eyben, Felix Weninger, Nicolas Lehment, Björn Schuller, and Gerhard Rigoll. 2014. Affective video retrieval: Violence detection in Hollywood movies by large-scale segmental feature extraction. *PLOS ONE* 8, 12 (2014), 1–9.

[8] Jiyang Gao, Zhenheng Yang, Kan Chen, Chen Sun, and Ram Nevatia. 2017. Turn tap: Temporal unit regression network for temporal action proposals. In *ICCV*.

[9] Sanjay Goswami, Satrajit Nandi, and Sucheta Chatterjee. 2019. Sentiment analysis based potential customer base identification in social media. In *Contemporary Advances in Innovative and Applicable Information Technology*.

[10] Lili Guo, Longbiao Wang, Chenglin Xu, Jianwu Dang, Eng Siong Chng, and Haizhou Li. 2021. Representation Learning with Spectro-Temporal-Channel Attention for Speech Emotion Recognition. In *ICASSP*.

[11] Fa-Ting Hong, Jia-Chang Feng, Dan Xu, Ying Shan, and Wei-Shi Zheng. 2021. Cross-modal Consensus Network for Weakly Supervised Temporal Action Localization. In *ACM MM*.

[12] He-Yen Hsieh, Ding-Jie Chen, and Tyng-Luh Liu. 2022. Contextual Proposal Network for Action Localization. In *WACV*.

[13] Linjiang Huang, Liang Wang, and Hongsheng Li. 2021. Foreground-Action Consistency Network for Weakly Supervised Temporal Action Localization. In *ICCV*.

[14] Linjiang Huang, Liang Wang, and Hongsheng Li. 2022. Multi-Modality Self-Distillation for Weakly Supervised Temporal Action Localization. *TIP* 31 (2022), 1504–1519.

[15] Haroon Idrees, Amir R Zamir, Yu-Gang Jiang, Alex Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. 2017. The THUMOS challenge on action recognition for videos "in the wild". *CVIU* 155 (2017), 1–23.

[16] Go Irie, Takashi Satou, Akira Kojima, Toshihiko Yamasaki, and Kiyoharu Aizawa. 2010. Affective audio-visual words and latent topic driving model for realizing movie affective scene classification. *TMM* 12, 6 (2010), 523–535.

[17] Haeng-Jin Jang, Jaemoon Sim, Yonnim Lee, and Ohbyung Kwon. 2013. Deep sentiment analysis: Mining the causality between personality-value-attitude for analyzing business ads in social media. *Expert Systems with applications* 40, 18 (2013), 7492–7503.

[18] Yuan Ji, Xu Jia, Huchuan Lu, and Xiang Ruan. 2021. Weakly-Supervised Temporal Action Localization via Cross-Stream Collaborative Learning. In *ACM MM*.

[19] Yu-Gang Jiang, Baohan Xu, and Xiangyang Xue. 2014. Predicting emotions in user-generated videos. In *AAAI*.

[20] Brendan Jou, Subhabrata Bhattacharya, and Shih-Fu Chang. 2014. Predicting viewer perceived emotions in animated GIFs. In *ACM MM*.

[21] Chen Ju, Peisen Zhao, Siheng Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. 2021. Divide and Conquer for Single-Frame Temporal Action Localization. In *ICCV*.

[22] Hang-Bong Kang. 2003. Affective content detection using HMMs. In *ACM MM*.

[23] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[24] Qiuqiang Kong, Yong Xu, Wenwu Wang, and Mark D Plumbley. 2017. A joint detection-classification model for audio tagging of weakly labelled data. In *ICASSP*.

[25] Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. 2019. Context-aware emotion recognition networks. In *ICCV*.

[26] Jun-Tae Lee, Sungrack Yun, and Mihir Jain. 2022. Leaky Gated Cross-Attention for Weakly Supervised Multi-Modal Temporal Action Localization. In *WACV*.

[27] Pilhyeon Lee and Hyeran Byun. 2021. Learning Action Completeness from Points for Weakly-supervised Temporal Action Localization. In *CVPR*.

[28] Pilhyeon Lee, Jinglu Wang, Yan Lu, and Hyeran Byun. 2021. Weakly-supervised temporal action localization by uncertainty modeling. In *AAAI*.

[29] Chuming Lin, Chengming Xu, Donghao Luo, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yanwei Fu. 2021. Learning salient boundary feature for anchor-free temporal action localization. In *CVPR*.

[30] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal Loss for Dense Object Detection. In *ICCV*.

[31] Jiaxing Liu, Sen Chen, Longbiao Wang, Zhilei Liu, Yahui Fu, Lili Guo, and Jianwu Dang. 2021. Multimodal Emotion Recognition with Capsule Graph Convolutional Based Representation Fusion. In *ICASSP*.

[32] Fuchen Long, Ting Yao, Zhaofan Qiu, Xinmei Tian, Jiebo Luo, and Tao Mei. 2019. Gaussian temporal awareness networks for action localization. In *CVPR*.

[33] Zhekun Luo, Devin Guillory, Baifeng Shi, Wei Ke, Fang Wan, Trevor Darrell, and Huijuan Xu. 2020. Weakly-supervised action localization with expectation-maximization multi-instance learning. In *ECCV*.

[34] Fengmao Lv, Xiang Chen, Yanyong Huang, Lixin Duan, and Guosheng Lin. 2021. Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences. In *CVPR*.

[35] Fan Ma, Linchao Zhu, Yi Yang, Shengxin Zha, Gourab Kundu, Matt Feiszli, and Zheng Shou. 2020. Sf-net: Single-frame supervision for temporal action localization. In *ECCV*.

[36] Kyle Min and Jason J Corso. 2020. Adversarial background-aware loss for weakly-supervised temporal activity localization. In *ECCV*.

[37] Trisha Mittal, Puneet Mathur, Aniket Bera, and Dinesh Manocha. 2021. Affect2mm: Affective analysis of multimedia content using emotion causality. In *CVPR*.

[38] Davide Moltisanti, Sanja Fidler, and Dima Damen. 2019. Action recognition from single timestamp supervision in untrimmed videos. In *CVPR*.

[39] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. 2019. When does label smoothing help?. In *NeurIPS*.

[40] Sanath Narayan, Hisham Cholakkal, Fahad Shahbaz Khan, and Ling Shao. 2019. 3c-net: Category count and center loss for weakly-supervised action localization. In *ICCV*.

[41] Phuc Nguyen, Ting Liu, Gautam Prasad, and Bohyung Han. 2018. Weakly supervised action localization by sparse temporal pooling network. In *CVPR*.

[42] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

[43] Phuong Pham, Juncheng Li, Joseph Szurley, and Samarjit Das. 2018. Eventness: Object detection on spectrograms for temporal localization of audio events. In *ICASSP*.

[44] Fan Qi, Xiaoshan Yang, and Changsheng Xu. 2021. Emotion Knowledge Driven Video Highlight Detection. *TMM* 23 (2021), 3999–4013.

[45] Sujata Rani and Parteek Kumar. 2017. A sentiment analysis system to improve teaching and learning. *Computer* 50, 5 (2017), 36–43.

[46] Julien Schroeter, Kirill Sidorov, and David Marshall. 2019. Weakly-supervised temporal localization via occurrence count learning. In *ICML*.

[47] Christian Schulze, Dominik Henter, Damian Borth, and Andreas Dengel. 2014. Automatic detection of CSA media by multi-modal feature fusion for law enforcement support. In *ICMR*.

[48] Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. 2017. A survey of multimodal sentiment analysis. *Image and Vision Computing* 65 (2017), 3–14.

[49] Statista. 2020. Hours of video uploaded to YouTube every minute as of February 2020. https://www.statista.com/statistics/259477/hours-of-video-uploaded-to-youtube-every-minute/. [Online].

[50] Chen Sun, Sanketh Shetty, Rahul Sukthankar, and Ram Nevatia. 2015. Temporal localization of fine-grained actions in videos by domain transfer from web images. In *ACM MM*.

[51] Jiajun Tang, Jin Xia, Xinzhi Mu, Bo Pang, and Cewu Lu. 2020. Asynchronous interaction aggregation for action detection. In *ECCV*.

[52] Paul Vicol, Makarand Tapaswi, Lluis Castrejon, and Sanja Fidler. 2018. MovieGraphs: Towards Understanding Human-Centric Situations from Videos. In *CVPR*.

[53] Paulo Vitorino, Sandra Avila, Mauricio Perez, and Anderson Rocha. 2018. Leveraging deep neural networks to fight child pornography in the age of social media. *Journal of Visual Communication and Image Representation* 50 (2018), 303–313.

[54] Hee Lin Wang and Loong-Fah Cheong. 2006. Affective understanding in film. *TCSVT* 16, 6 (2006), 689–704.

[55] Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. 2017. Untrimmednets for weakly supervised action recognition and detection. In *CVPR*.

[56] Shangfei Wang and Qiang Ji. 2015. Video affective content analysis: a survey of state-of-the-art methods. *TAC* 6, 4 (2015), 410–430.

[57] Baohan Xu, Yanwei Fu, Yu-Gang Jiang, Boyang Li, and Leonid Sigal. 2018. Heterogeneous Knowledge Transfer in Video Emotion Recognition, Attribution and Summarization. *TAC* 9, 2 (2018), 255–270.

[58] Huijuan Xu, Abir Das, and Kate Saenko. 2017. R-c3d: Region convolutional 3d network for temporal activity detection. In *ICCV*.

[59] Jufeng Yang, Dongyu She, and Ming Sun. 2017. Joint Image Emotion Classification and Distribution Learning via Deep Convolutional Neural Network.. In *IJCAI*.

[60] Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. 2020. Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *ACL*.

[61] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor Fusion Network for Multimodal Sentiment Analysis. In *EMNLP*.

[62] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259* (2016).

[63] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *ACL*.
[64] Runhao Zeng, Wenbing Huang, Mingkui Tan, Yu Rong, Peilin Zhao, Junzhou Huang, and Chuang Gan. 2021. Graph Convolutional Module for Temporal Action Localization in Videos. *TPAMI* (2021).
[65] Yuanhao Zhai, Le Wang, Wei Tang, Qilin Zhang, Junsong Yuan, and Gang Hua. 2020. Two-stream consensus network for weakly-supervised temporal action localization. In *ECCV*.
[66] Can Zhang, Meng Cao, Dongming Yang, Jie Chen, and Yuexian Zou. 2021. CoLA: Weakly-Supervised Temporal Action Localization with Snippet Contrastive Learning. In *CVPR*.
[67] Haimin Zhang and Min Xu. 2018. Recognition of emotions in user-generated videos with kernelized features. *TMM* 20, 10 (2018), 2824–2835.
[68] Shiqing Zhang, Shiliang Zhang, Tiejun Huang, Wen Gao, and Qi Tian. 2017. Learning affective features with a hybrid deep model for audio–visual emotion recognition. *TCSVT* 28, 10 (2017), 3030–3043.
[69] Hang Zhao, Antonio Torralba, Lorenzo Torresani, and Zhicheng Yan. 2019. Hacs: Human action clips and segments dataset for recognition and temporal localization. In *ICCV*.
[70] Sicheng Zhao, Guoli Jia, Jufeng Yang, Guiguang Ding, and Kurt Keutzer. 2021. Emotion Recognition From Multiple Modalities: Fundamentals and methodologies. *SPM* 38, 6 (2021), 59–73.
[71] Sicheng Zhao, Yunsheng Ma, Yang Gu, Jufeng Yang, Tengfei Xing, Pengfei Xu, Runbo Hu, Hua Chai, and Kurt Keutzer. 2020. An End-to-End Visual-Audio Attention Network for Emotion Recognition in User-Generated Videos. In *AAAI*.
[72] Tao Zhao, Junwei Han, Le Yang, Binglu Wang, and Dingwen Zhang. 2021. Soda: Weakly supervised temporal action localization based on astute background response and self-distillation learning. *IJCV* 129, 8 (2021), 2474–2498.
[73] Jia-Xing Zhong, Nannan Li, Weijie Kong, Tao Zhang, Thomas H Li, and Ge Li. 2018. Step-by-step erasion, one-by-one collection: a weakly supervised temporal action detector. In *ACM MM*.
[74] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million image database for scene recognition. *TPAMI* 40, 6 (2017), 1452–1464.