Weakly Supervised Video Emotion Detection and Prediction via Cross-Modal Temporal Erasing Network

Zhicheng Zhang Lijuan Wang Jufeng Yang†
TMCC, College of Computer Science, Nankai University, China
gloryzzc6@sina.com, 13693225189@163.com, yangjufeng@nankai.edu.cn

Abstract

Automatically predicting the emotions of user-generated videos (UGVs) receives increasing interest recently. However, existing methods mainly focus on a few key visual frames, which may limit their capacity to encode the context that depicts the intended emotions. To tackle that, in this paper, we propose a cross-modal temporal erasing network that locates not only keyframes but also context and audio-related information in a weakly-supervised manner. In specific, we first leverage the intra- and inter-modal relationship among different segments to accurately select keyframes. Then, we iteratively erase keyframes to encourage the model to concentrate on the contexts that include complementary information. Extensive experiments on three challenging video emotion benchmarks demonstrate that our method performs favorably against state-of-the-art approaches. The code is released on https://github.com/nku-zhichengzhang/WECL.

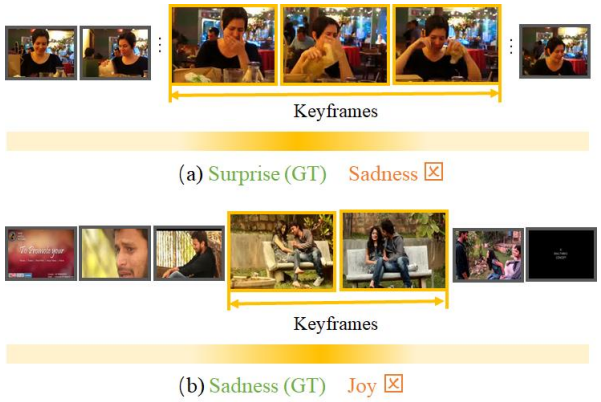

Figure 1. Illustration of the keyframes with larger boxes that are detected by off-the-shelf method [65] on the Ekman-6 dataset [54]. Note that the deeper color represents the higher impact on overall video emotion, and "GT" represents the category label from the ground truth. The texts with orange color are the categorical predicting results.

1. Introduction

Emotion analysis in user-generated videos (UGVs) has attracted much attention since a growing number of people tend to express their views on social networks [20, 32, 36]. Automatic predictions of video emotions [52, 54] can potentially be applied in various areas like online content filtering [1], attitude recognition [32], and customer behavior analysis [39]. Emotions evoked in UGVs usually depend on multiple perspectives, such as actions, events, and objects, where different frames in a UGV may contribute unequally to conveying emotions.

Existing methods in this field mainly focus on extracting keyframes from visual content, assuming that these frames hold the dominant information for the intended emotions in videos. For example, Tu *et al*. [12] introduce an attribution network to locate keyframes with temporal annotations, which are more precise than video-level labeling and lead to better performance for video emotion recognition. Yet, annotating the emotional labels frame-by-frame is labor-sensitive and time-consuming [63]. Zhao *et al*. [65] further present the visual-audio network (VAANet) conducting three types of attention to automatically discover the discriminative keyframes, which makes it state-of-the-art.

However, the selected "keyframes" may fail to represent the intended emotions exactly due to the inherent characteristics of human emotions, *i.e*., subjectivity and ambiguity [49, 57, 66]. As illustrated in Figure 1 (a), a woman receives a gift and moves to cry. The video-level emotion category is labeled as 'surprise'. We could observe that VAANet [65] gives the most attention to the keyframes (*i.e*., "crying" frames in the larger boxes) while ignoring the context and leading to the wrong prediction. Furthermore, in Figure 1 (b), a man sees his beloved girl talking with another man happily, which makes him feel sad. However, VAANet only focuses on frames about chatting and categorizes the emotion of this video as 'joy'. Therefore,

† Corresponding author.

keyframes may lead to limited prediction results. Although the detected keyframes directly convey emotions in most videos, some other information that contains the necessary context should not be ignored. This is because the contextual frames could not only provide complementary information for understanding emotions in UGVs (especially for the cases where keyframes are hard to be distinguished), but also make the model more robust since more cues are considered to recognize emotions rather than the dominant information.

To address these problems, we propose a novel cross-modal temporal erasing network, which enables the model to be aware of context for recognizing emotions. Our proposed method mainly contains two crucial modules: temporal correlation learning module, and temporal erasing module. First, we extract the visual feature together with the audio one from each equal-length segment derived from the video. Second, the temporal correlation learning module is introduced to fully discover comprehensive implicit correspondences among different segments across audio and visual modal. Then, keyframes are selected by considering the correspondences with other frames in a weakly-supervised manner, where only the video-level class label is used. Finally, the temporal erasing module iteratively erases the most dominant visual and audio information to train with difficult samples online, that encouraging the model to detect more complementary information from context instead of the most dominant ones.

Our contributions can be summarized as follows: 1) We introduce a weakly-supervised network to exploit keyframes and the necessary context in a unified CNN framework, which encourages model to extract features from multiple discriminative parts and learn better representation for video emotion analysis. 2) We exploit intra- and inter-modal relationships to provide frame-level localized information only with video-level annotation, with which the model consolidates both holistic and local representations for affective computing. We demonstrate the advantages of the above contributions with extensive experiments. Our method achieves state-of-the-art on three video emotion datasets.

2. Related Work

In this section, we introduce some existing works that are related to our work. We clearly divide them into two groups: a) Video Emotion Recognition, and b) Weakly-Supervised Learning in Videos.

2.1. Video Emotion Recognition

Existing work on video emotion prediction mainly depends on the emotion models of categorical and dimensional assumptions [64]. In the categorical assumption [19, 24, 27, 28, 40, 46, 56], emotions are described by a fixed number of emotional classes. Take an instance, Ekman's psychological research [11] points out that there are six basic emotions across cultures and countries: anger, disgust, fear, joy, sadness, and surprise, which existed universally. In contrast, dimensional assumption [38] usually represents emotions in a continuous way. The most commonly used dimensional spaces are 3D and 2D Cartesian spaces (*e.g.*, Valence Arousal Dominance (VAD) and activity-temperature-weight). Compared to dimensional models, the categorical model can be easier understood by users attributed to straightforwardness. In this paper, we adopt the categorical approach to predict emotions in UGVs.

Early works [7, 20, 54] mainly focus on designing representative features, aiming to recognize the highly-abstract emotion. Jiang *et al*. [20] introduce numerous low-level and mid-level features, including OBank [23] and SBank [4] to recognize emotions. Chen *et al*. [7] propose to employ existing detectors to learn various high-level semantic features. However, it is not applicable in real-world applications where their required auxiliary data is missed. Sikka *et al*. [43] simply combine multiple visual descriptors with para-linguistic audio features for multi-modal emotion classification of video clips.

Recently, compared to hand-crafted features, deep features have demonstrated their superior representation ability to predict emotions in videos [29, 61, 65]. Zhang *et al*. [61] extract frame-level deep features, and then used discrete Fourier transform to obtain kernelized features for recognizing emotions. M3ER [29] is a learning-based fusion method, which aims at emphasizing more reliable features and suppressing others. VAANet [65] proposes the first deep framework to recognize emotions in UGVs, which includes three attention modules to automatically capture the most discriminative keyframes and extract robust affective representation. Although the above methods achieve significant improvement in emotion recognition in UGV, they mainly focus on the most dominant information and neglect the necessary contextual information.

2.2. Weakly-Supervised Learning in Videos

Intuitively, our work targets selecting a set of segments containing rich, comprehensive information to make video-level emotion recognition accurate and unambiguous. Our method is closely related to temporal action location [13, 17, 25, 26, 60]. The objectiveness of temporal action location is to localize the beginning and end of each action within an offered video and recognize the action category. Although fully-supervised methods [41, 58] have achieved remarkable success in video action analysis, weakly-supervised researches [13, 17, 25] also become popular for action analysis in videos with less annotation burden.

The supervisory information includes movie scripts [9,

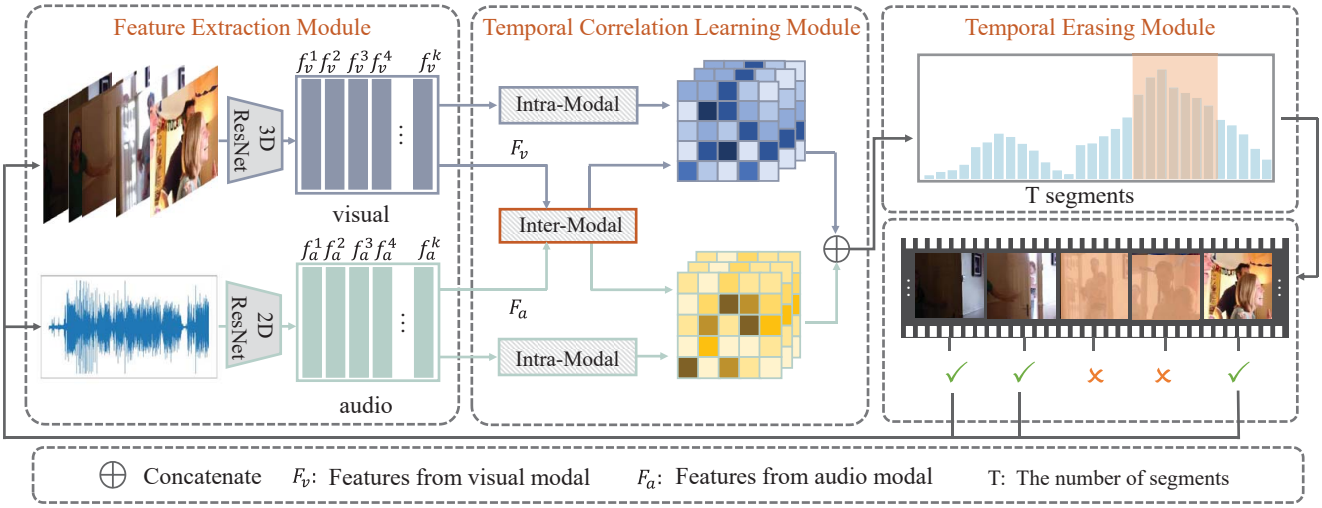

Figure 2. Illustration of our pipeline. Given a video, it naturally contains two types of modal (*i.e.*, visual and audio). First, we employ 3D ResNet-101 to extract the features of the visual stream and utilize 2D ResNet-50 to extract features from the audio stream. Second, the temporal correlation learning module is introduced to model the intra- and inter-modal relationship of two modalities. Then, the temporal erasing module erases the most dominant visual and audio information to generate difficult training samples online. Finally, the training samples are input into the model to find out more complementary information.

21], action lists [3, 16], or video-level class label [30, 34, 44, 45, 50], *etc*. Duchenne *et al*. [9] propose to employ scripts to localize action categories. While the supervised information of movie scripts can only apply to movie videos, our method analyzes emotion in videos. Temporally ordered action lists are another type of supervisory information for action analysis in videos. For example, Bojanowski *et al*. [3] propose to weakly supervise label action instances in videos via the discriminative clustering method. Huang *et al*. [16] introduce the extended framework of Connections Temporal Classification [14] to weakly-supervised action labels from speech recognition. Different from these methods that contain order information about the containing action instances, the studied UGVs have no specific order information and only own video-level emotional categories.

Video-level category annotation is the most widely-studied weak supervision and provides the least label information. Sun *et al*. [45] present the first work to introduce only video-level categorical annotations for weakly-supervised labeling. Singh *et al*. [44] propose a framework based on an augmentation strategy that hides patches and encourages the model to learn the most discriminative segments. Wang *et al*. [50] design a selection module to localize temporal action segments in videos. STPN [30] is introduced to use a sparse loss function to help models choose temporal action segments. Then, Paul *et al*. [34] propose to train a weakly-supervised network by jointly optimizing two complimentary losses. However, these methods are all based on a single modality to select action segments. UGVs usually evoke emotions from multi-modal information. Besides, UGVs usually may contain multiple emotions but only contains one dominant emotion. Different from the above works that try to locate all types of actions in the video; our task is to locate frames that are related to the most dominant emotion.

3. Methodology

3.1. Visual-Audio Representation Extraction

Visual Representation Extraction: To obtain visual representations from user-generated videos, we follow [50, 65] to extract visual features. First, given a video, we split it into T equal-length segments and randomly select k successive frames from every segment. Second, we adopt 3D ResNet-101 [15] to extract each segment features in every video. Then, it takes the T snippet as the input and processes them independently. Therefore, for the given video l, the output is a set of segment features that can be represented as $F_v(l) = f_v^1(l), f_v^2(l), f_v^3(l), ..., f_v^T(l)$. For each segment-level features $f_v^i(l)$, the $f_v^i(l) \in R^{H \times W \times C}$ and H, W indicate height and width of feature map, respectively. Besides, C is the dimension of the feature.

Audio Representation Extraction: The audio stream can be considered auxiliary information for the visual stream. We follow the work [65] to use the most commonly-used audio feature description, *i.e.*, the Mel-Frequency Cepstral Coefficients (MFCC). Given a video, we can obtain a successive discriminator for the audio stream through MFCC. We separate the successive discriminator into T segments, which are the same as the visual

stream. For each descriptor $F_a(l)$ in the video, we input it into 2D ResNet-50 and obtain its representation $F_a(l) = f_a^1(l), f_a^2(l), f_a^3(l), ..., f_a^T(l)$. Then, the $F_a^i(l)$ consists of a 3-dimensional matrix $R_A^{H^{'} \times W^{'} \times C^{'}}$, where $H^{'}, W^{'}$ indicate the height and width of the audio feature map, respectively. The $C^{'}$ is regarded as the dimension of the feature. Note that C and $C^{'}$ have the same values.

3.2. Temporal Correlation Learning

In Section 3.1, we have extracted the video feature F_v and audio feature F_a via pre-trained 3D ResNet-101 and 2D ResNet-50, respectively. Then, the features F_v and F_a reshape into $F_v \in R^{T \times C}$ and $F_a \in R^{T \times C^{'}}$ after spatial average pooling. As discussed in [51], the 3D CNNs can not model the correlation between different segments in each video directly, restricted by its receptive field. Meanwhile, we use 2D CNNs to extract audio features on audio patches, which cannot also model this correlation. To better locate important segments in video and audio sequences, we need to consider the correlation of different segments. Motivated by this, we propose to enhance the learned video features from the following two aspects. First, we model the intra-modal correlation among different segments. Second, we introduce the inter-modal attention module to learn more complementary information about different frames from multi-modal information.

Intra-Modal Relation Modeling: Given T segments video and audio sequence, we obtain the video features $F_v \in R^{T \times C}$ and audio feature $F_a \in R^{T \times C^{'}}$ (we denote F for convenience), where C and $C^{'}$ are the dimension of extracted features. Inspired by the classical non-local attention mechanisms [5, 48] in computer vision, we develop the intra-modal attention module to model the long-range dependencies between different frames in each modal. In specific, we first embed the features F into three subspaces through three linear project functions θ, ϕ, g:

$$Q = \theta(\mathbf{F}), K = \phi(\mathbf{F}), V = g(\mathbf{F}), \tag{1}$$

where $\mathbf{Q} \in R^{T \times C}$, $\mathbf{K} \in R^{T \times C}$ and $\mathbf{V} \in R^{T \times C}$ means the key, query and value features. The temporal correlation of different segments in F is encoded via dot product similarity between the query and key features, which can be formulated as the following equation:

$$S(Q, K) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}}), \tag{2}$$

where we apply softmax normalization on each row and the similarity matrix $S(Q, K) \in R^{T \times T}$ encoded the correlation between each query segment and all segments. In order to enhance the feature of each segment with other temporal segments, we fuse the value features by the temporal correlation weight S, which can be implemented as a matrix multiply:

$$Z = S(Q, K)V, \tag{3}$$

where $Z \in R^{T \times C}$ have the same dimension with input feature F. In order to incorporate the intra-modal temporal module into the pre-trained network and ease its optimization, we add a residual connection as a short path between the input feature and the enhanced feature:

$$F' = F + W_z Z, \tag{4}$$

where W_z is the learnable parameter that controls the importance of intra-modal temporal fusion. We employ $F_a^{'}$ and $F_v^{'}$ to represent the enhanced features for F_a and F_v, respectively. Overall, the intra-modal temporal module exploits the pair-wise correlation $< f_i, f_j >$ of video and audio features, where i and j are the indexes of segments. Such modeling exploits the long-range dependency between each segment, which complements the short-range information encoded by pre-trained 2D ResNet-50 and 3D ResNet-101.

Inter-Modal Relation Modeling The intra-modal temporal fusion can enhance the extracted feature by capturing their global temporal dependence. However, each modal has complementary information to other modalities, which is also beneficial to learn the relationship between different frames. Motivated by this, we develop inter-modal attention to learn more robust representations. We denote $S_{v \leftarrow a}(Q_v, K_a)$ and $S_{a \leftarrow v}(Q_a, K_v)$ in Eq.(2) as the temporal correlation matrix extracted between video and audio features, respectively. In intra-modal temporal fusion, the correlation matrix is used to guide the fusion within the original modal, *i.e.*, the enhanced video feature $F_v^{'}$ is a weighted sum of visual value feature F_v. In some cases, the temporal correlation between different segments cannot be well represented by a single modality. For example, in the evening, the dark environment can not convey meaningful visual information. We need audio features to guide the visual temporal fusion and vice versa. To be specific, in the inter-modal attention module, we use the features of the audio modal F_a to guide the calculation of visual temporal correlation. Meantime, the audio temporal correlation is guided by the visual features F_v. The Eq.(3) in intra-modal temporal fusion can be formulated as follows:

$$U_v = S_{v \leftarrow a}(Q_v, K_a)V_a, \quad U_a = S_{a \leftarrow v}(Q_a, K_v)V_v, \tag{5}$$

where U_v and U_a mean inter-modal enhanced features for the audio and visual modal, respectively. We also add a residual connection to input features to ease optimization, which is formulated as:

$$F_v^{''} = F_v + W_v U_v, \quad F_a^{''} = F_a + W_a U_a, \tag{6}$$

where W_v and W_a are the learnable parameters.

After extracting the enhanced features from the visual and audio streams, we concentrate them together along with their dimension. Overall, the inter-modal attention module exploits the pair-wise correlation and guides the fusion by cross-modal temporal correlation. The inter-modal attention module can make two modalities complement each other. With the proposed intra- and inter-modal attention modules, we can learn more representative features for recognizing emotions in UGVs.

3.3. Temporal Erasing Module

As mentioned above, we first introduce how to use intra-modal attention to capture long-range dependencies between frames in each modal. The inter-modal attention module aims to model the relationship between audio and visual modal, which is beneficial to guide the fusion of two modalities. Within the above two modules, we could better learn the relationship between the frames in each UGV. According to the existing work [47, 65], we could know that different frames contribute differently to video emotion recognition. Instead of keyframes that directly evoke emotions, some other frames that contain background or context also play a very important role in understanding the emotion of UGVs conveyed. However, it is usually hard to locate contextual frames since they are usually less significant. Inspired by advanced work [53, 62] that erase key regions to find the complementary ones, we introduce a simple yet effective way to find more complementary frames for perceiving highly-abstract emotions in UGVs, *i.e.*, temporal erasing module. Our temporal erasing module erases the dominant information guided by attention weights as important indicators, which encourages the model to investigate both complementary evidence and the dominant one.

In detail, the model sufficiently knows the relationship between different frames in each UGV from the aforementioned attention modules. Then, we apply the temporal attention module following [65] to automatically find out important segments. The temporal attention module for visual A_v^T and audio A_a^T stream is defined as:

$$\begin{aligned} A_v^T &= \mathrm{Relu}(W_1(W_2(F_v^{'} + F_v^{''})^\top)^\top), \\ A_a^T &= \mathrm{Relu}(W_1^{'}(W_2^{'}(F_a^{'} + F_a^{''})^\top)^\top), \end{aligned} \tag{7}$$

where W_1, W_2, $W_1^{'}$, $W_2^{'}$ represents four learnable parameter matrices and $\top$ represents the transpose of a matrix. Then, we normalize A_v^T and A_a^T attention maps with the following equation:

$$\begin{aligned} A_v^* &= \frac{A_v^T - \min(A_v^T)}{\max(A_v^T) - \min(A_v^T)}, \\ A_a^* &= \frac{A_a^T - \min(A_a^T)}{\max(A_a^T) - \min(A_a^T)}. \end{aligned} \tag{8}$$

With the temporal attention maps A_v^* and A_a^*, we can find the keyframes and erase them, in order to drive our model to seek more complementary information. The erasing mask is defined as follows:

$$E^T = \begin{cases} 1, & if\ A_v^* \geq \theta\ \mathrm{and} A_a^*\ \geq \theta, \\ 0, & \mathrm{otherwise}. \end{cases} \tag{9}$$

Where θ is a hyper-parameter that can be set manually. According to the erasing mask, we erase some segments, and the left segments form a new UGV. Then, the left segments are input into the network again, in order to force the network to find more complementary information for video emotion recognition. Note that this module is independent of the backbone architecture, and can be applied via any attention-based structures without adding extra model parameters or complexity.

3.4. Optimizing Process

As mentioned above, we can detect keyframes from the temporal attention module in a weakly-supervised manner. Then, the left frames except keyframes may be contextual frames. To learn more representative features, the loss function of our proposed method contains three parts, *e.g.*, the original video, keyframes, and the left frames. Given a training set $\{x_i, y_i\}_{i=1}^N$, where x_i indicates the i^{th} video and $y_i \in \{1, 2, ..., M\}$ is the single class label. M is the number of classes. We employ the cross-entropy loss function to optimize our proposed method, which is defined as follows:

$$\ell_{ce}(x, y) = -\frac{1}{N}\left[\sum_{i=1}^{N}\sum_{j=1}^{M}(y_i = j)\ln p^j(x_i)\right]. \tag{10}$$

The $p^j(x_i)$ represents the probability that the input video x_i belongs to class j, which is formulated as follows:

$$p^j(x_i) = \frac{\exp(a_i^j)}{\sum_{k=1}^{M}\exp(a_i^k)}, \tag{11}$$

where $\{a_i^k | k = 1, 2, ..., M\}$ are the activation values of units in the last fully-connected layer for the input video x_i. The overall loss function ℓ that is used to optimize our proposed method is defined as:

$$\ell = \ell_{ce}(x_o, y) + \ell_{ce}(x_k, y) + \ell_{ce}(x_l, y), \tag{12}$$

where x_o, x_l, and x_l represent the original video, keyframes, and the left frames, respectively. Therefore, this work contains two tasks: video-level emotion recognition (i.e., the summation of three cross-entropy losses) and keyframes detection. The former falls into a fully-supervised task since we employ video-level emotional categories to optimize the network, while the latter is trained in a weakly-supervised manner.

Table 1. Comparison with state-of-the-art methods about video emotion recognition. Classification accuracy (%) on the testing set of VideoEmotion-8, Ekman-6, and CAER datasets. Note that "Video-8" means the VideoEmotion-8 dataset. "N/A" means the experimental results are not available.

Method	[18]	[7]	[61]	[20]	[33]	[55]	[54]	[31]	[37]	[2]	[59]	[65]	**Ours**
Modal	V	V	V	V+A	V+A	V+A	V+A	V+A	V+A	V+A	V+A	V+A	V+A
Video-8	39.3%	50.6%	52.5%	46.1%	51.1%	51.4%	52.6%	53.3%	53.3%	53.7%	54.2%	54.5%	**57.3%**
Ekman-6	41.3%	51.8%	54.4%	N/A	N/A	51.2%	55.6%	N/A	57.3%	54.2%	54.3%	55.3%	**58.2%**
CAER	52.1%	N/A	N/A	N/A	N/A	N/A	77.9%	N/A	N/A	77.3%	78.7%	78.3%	**80.1%**

4. Experiment

4.1. Datasets

VideoEmotion-8 dataset [20] consists of 1,101 UGVs scrawled from two video-sharing websites of Youtube and Flicker, which contains at least 100 videos in each emotional category. According to the emotion model of Plutchik Wheel, each video is labeled by eight emotion classes [35].
Ekman-6 dataset [54] is similarly collected from social websites. Each video is labeled with one kind of emotion that is based on Ekman's psychology research [10]. As a result, there are 1,637 videos.
CAER dataset [22] is collected from TV shows, and the length of sequences averages 90. Each video is manually labeled with six basic emotions, which is the same as the Ekman-6 dataset. The overall number of clips is 13,201.

4.2. Implementation Details

Following [65], we use 2D ResNet-50 and 3D ResNet-101 as the audio and visual feature extractor, respectively. We employ the audio and visual extractor pre-trained on ImageNet [8] and Kinetics-400 [6]. For the visual stream, each of the videos is divided into 10 segments and samples 16 successive frames in each segment. Video frames are resized to the size of 112×112, and audio MFCC features are resized to 224×224. The minimum, maximum, average, and variance of lengths in seconds of the two employed datasets are 1,505, 107, 45.32 and 3, 635, 112, 60.26 on Ekman-6 and VideoEmotion-8, respectively. We use the random center crop as data augmentation for the visual stream to reduce over-fitting. Our framework is optimized by the mini-batch SGD, where the size is 32. We fixed the feature extractor. We train the parameter of the attention module and linear classifier by a learning rate of 2e-4. The dataset splitting is following the same configuration as [20, 61, 65]. On the VideoEmotion-8 dataset, we conduct ten runs of experiments. In each run, the proportion of training and testing is 2:1. In the Ekman-6 dataset, the proportion of training and testing is 1:1, thus confirming the generalization of evaluation. In the CAER dataset, the proportion of training, validation, and testing is 7:1:2. We use the mean accuracy among ten splits as the overall metric for evaluation. During inference, we use the output based on the original video, *i.e.*, predicted by x_o. We conduct classification by a last fully-connected layer designed with 4,096 neurons. We use Pytorch to implement our method. All of our experiments are performed on two 3090 GPUs.

4.3. Video Emotion Prediction

Comparison with SOTAs: Table 1 compares our proposed method to existing advanced approaches on three video emotion datasets (*i.e.*, Ekman-6, VideoEmotion-8, and CAER). Some experimental results are not available since the code is not released to the public. Our proposed approach achieves competitive performance against state-of-the-art methods. Our method obtains 2.8%, 2.9%, and 1.8% improvement on the VideoEmotion-8, Ekman-6, and CAER datasets, respectively. Experimental results verify that ours learns robust affective representation for video emotion recognition with the help of temporal correlation learning and temporal erasing module. We compare several previous successful video emotion classification methods on three datasets, including conventional video analysis, uni-modal emotion recognition methods, and multi-modal emotion recognition methods.

As illustrated in Table 1, we could observe that all methods employ visual features to recognize emotion since they are more direct to evoke emotions in videos. It demonstrates that visual features are important to learn discriminative representation for video emotion recognition. Although only using the audio stream can not gain better performance for video analysis as shown in Table 2, an interesting observation is that audio can improve the performance of models when combined with the visual modal. This is because audio can be regarded as auxiliary information to assist models to decide the better result. Therefore, we could observe that methods using multi-modal features achieve better results. However, previous methods [20, 33, 54] simply concatenate audio and visual features without considering the relationship between them, which leads to limited improvement. Existing work [2, 31, 59, 65] employ more complex and adaptive multi-modal feature fusion methods and improve the performance. Our method also employs multi-modal features and leverages the intra- and inter-modal re-

Table 2. Ablation study of three modules for the video emotion datasets. Our proposed temporal correlation learning includes the intra- and inter-modal attention modules. Another main module is the temporal erasing module. Note that the "intra", "inter", "erasing" represents the intra-modal attention module, the inter-modal attention module, and temporal erasing module, respectively.

Modal	Modules			Datasets	
	Intra	Inter	Erasing	Video-8	Ekman-6
Audio	✗	✗	✗	40.5%	36.5%
	✓	✗	✗	41.1%	38.0%
	✓	✗	✓	42.0%	40.1%
Visual	✗	✗	✗	51.3%	51.5%
	✓	✗	✗	52.6%	53.2%
	✓	✗	✓	53.1%	54.8%
Audio & Visual	✗	✗	✗	52.3%	52.5%
	✓	✗	✗	53.1%	54.7%
	✗	✓	✗	53.5%	54.4%
	✓	✓	✗	54.1%	55.2%
	✓	✗	✓	55.2%	56.1%
	✗	✓	✓	55.9%	56.5%
	✓	✓	✓	57.3%	58.2%

lationship to further improve performance.

Hyper-parameter Analysis: We investigate the effect of θ in our proposed method as shown in the left panel of Figure 3, which is the threshold to select important segments. When the value of θ is zero, it represents all frames are important segments. Meanwhile, when its value is one, only the segment with the maximum attention weight is regarded as an important segment. As can be seen, with the increases of θ, our method boosts the performance against the baseline model ($\theta = 0$) without using the temporal erasing module. When θ increases from 0 to 0.7, the performance of prediction is boosted dramatically. Then, experiments show the performance is decreased when it increases from 0.7 to 1. It demonstrates that the frames that are related to the video-level emotional category mainly account for 30%. When the value is set higher than 0.7, the performance of the model is decreased. This is because θ influences the model optimization process. When the θ becomes higher, the model over-concentrates on the import segments and decreases its performance. It illustrates that context is essential to video emotion recognition.

4.4. Ablation Study

Effectiveness of intra-modal attention: From the results in Table 2, we find that intra-modal attention improves the performance of the model in all modalities. This verifies that modeling this module benefits the performance of video emotion analysis, which is consistent with previous works [51]. Therefore, modeling the relationship of intra-modal is meaningful, which indicates that this module is indispensable. Besides, we see a better performance of the model obtained from the visual stream than the one yielded from the audio stream. The reason for better results achieved by visual modality is that, visual content is the most straightforward media carrier for expressing emotions in the video. Intuitively, visual stimuli can largely arise human emotions such as surprise and fear.

Table 3. Comparison between our proposed method and existing SOTA methods on video emotion detection. We introduce mAP@tIoU for evaluation.

tIoU (α)	α= 0.5	α = 0.4	α = 0.3	α = 0.2	α = 0.1
[50]	13.9%	19.8%	26.2%	32.7%	35.8%
[67]	14.9%	21.2%	30.0%	35.7%	39.7%
[34]	18.9%	27.3%	37.8%	**45.3%**	50.6%
[42]	18.6%	26.1%	34.6%	41.8%	**53.5%**
[65]	19.7%	30.1%	38.7%	42.6%	49.7%
Ours	**20.1%**	**38.2%**	**40.1%**	43.5%	52.7%

Effectiveness of inter-modal attention: The inter-modal attention is designed to model the relationship across audio and visual modalities. Intuitively, the information between different modalities can complement each other, which can be helpful for deep feature embedding. Thus, the performance boost gained from the inter-modal attention module can be found in Table 2. This reveals that modeling the relationship between two modalities is beneficial. Although the experimental results of the audio stream are worse than the visual stream, the results after the fusion are even better.

Effectiveness of erasing module: The erasing module aims at preventing over-emphasizing the significant emotional context and encouraging the network to find out more complementary cues for video emotion analysis. We mainly conduct three types of experiments to show the effectiveness of this module. The first type is only using intra-modal attention to guide erasing. As shown in Table 2, the temporal erasing module not only improves the performance of the model whenever in uni-modal or multi-modal, which points out that this module is plug-and-play. The second type is purely employing inter-modal attention to guide the temporal erasing module. The third type is utilizing intra- and inter-modal attention simultaneously to direct this module. The erasing region is selected according to the attention weights. Thus, attention influences the effect of the temporal erasing module. As mentioned above, the fusion of intra- and inter-modal attention is the most effective for video emotion analysis. It is reasonable that the usage of three modules results in optimal results.

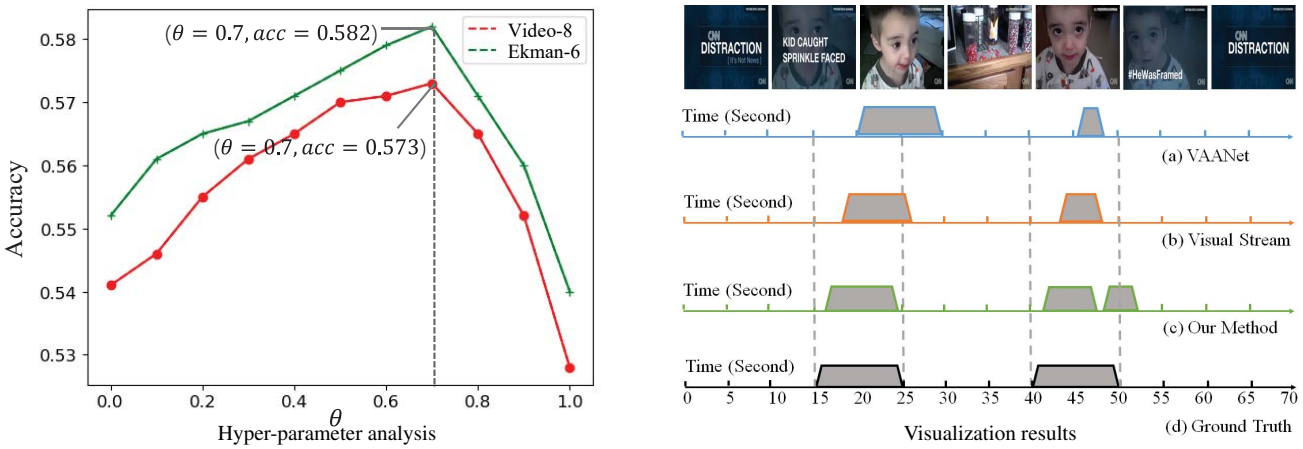

Figure 3. **Left**: Effect of the hyper-parameter θ on the Ekman-6, VideoEmotion-8 dataset, respectively. **Right**: Visualization of detected keyframes. A little boy is found eating snacks and shows a surprised expression. (a) is the prediction from VAANet. (b)-(c) are our predictions from the visual modality and overall modalities, respectively. (d) shows the ground truth of labeled keyframes from [12].

4.5. Video Emotion Detection

For the weakly-supervised detection, we follow the standard protocol that is based on mean average precision (mAP) at different temporal Intersection-over-Union (tIoU) thresholds. Given testing videos, our proposed method localizes a list of emotion segment predictions with the corresponding score for each prediction. Different from conventional methods that consider each prediction contains one action class and calculate mAP for each category. We consider the final video emotion category could be evoked by several types of emotion. The final detection result is regarded as the correct one only when its tIoU with the ground truth segment outperforms the evaluated threshold. We use the Ekman-6 dataset to verify the effectiveness of our method, in which the tIOU thresholds are $\{0.1, 0.2, 0.3, 0.4, 0.5\}$. We compare with five weakly-supervised video detection methods, as shown in Table 3. Both [50] and [65] employ the temporal attention module to locate keyframes. [65] is based on [50] that leverages temporal attention module to locate keyframes and further improve [50] the performance by adding audio features. [67] employs a step-by-step erasing strategy to find out segments that are related to the predominant class. It demonstrates that the audio is also helpful to predict video emotion. [34] considers temporal correlation among adjacent segments, while the correlation of long-range segments is not considered. To address this problem, [42] proposes a kind of contrastive loss to contrast the current segments with others. Compared with other methods, ours improves mean average precision, which illustrates the effectiveness of taking intra- and inter-modal information into consideration and further finding out more complementary information via temporal erasing module for locating emotional segments.

Visualization: The visualization is shown on the right side of Figure 3. The temporal emotion detection results are illustrated through the timelines. It illustrates that keyframes detected by our method are the most accurate and complete. That means the proposed method is better at mining more complementary information than others, attributed to the help of temporal correlation learning and temporal erasing module. With the multi-modal information, our proposed method focuses on more accurate segments, which demonstrates the multi-modal information is helpful in analyzing emotions in UGVs. Further, the temporal erasing module erases the most dominant information and encourages the model to find out more complementary information from context. Therefore, the detected keyframes become closer to the ground truth.

5. Conclusion

In this paper, we introduce to recognize and detect emotional segments from user-generated videos in a weakly-supervised manner. Our proposed method aims at preventing models from over-emphasizing the most significant emotional context and driving models to discover more complementary information. We design an intra- and inter-modal attention-guided erasing module that encourages the model to learn more complementary information, which consists of two crucial modules. We have systematically studied the effectiveness of our idea and performed extensive experiments to validate the proposed method.

6. Acknowledgments

This work was supported by the National Key Research and Development Program of China Grant (NO. 2018AAA0100400), Natural Science Foundation of Tianjin, China (NO.20JCJQJC00020), and Fundamental Research Funds for the Central Universities.

References

[1] Sharifa Alghowinem. A safer youtube kids: An extra layer of content filtering using automated multimodal analysis. In *Intelligent Systems and Applications*, 2019. 1

[2] John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. Gated multimodal units for information fusion. *ArXiv*, 2017. 6

[3] Piotr Bojanowski, Rémi Lajugie, Francis Bach, Ivan Laptev, Jean Ponce, Cordelia Schmid, and Josef Sivic. Weakly supervised action labeling in videos under ordering constraints. In *ECCV*, 2014. 3

[4] Damian Borth, Tao Chen, Rongrong Ji, and Shih-Fu Chang. Sentibank: large-scale ontology and classifiers for detecting sentiment and emotions in visual content. In *ACM MM*, 2013. 2

[5] Antoni Buades, Bartomeu Coll, and J-M Morel. A non-local algorithm for image denoising. In *CVPR*, 2005. 4

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

[7] Chen Chen, Zuxuan Wu, and Yu-Gang Jiang. Emotion in context: Deep semantic feature fusion for video emotion recognition. In *ACM MM*, 2016. 2, 6

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009. 6

[9] Olivier Duchenne, Ivan Laptev, Josef Sivic, Francis Bach, and Jean Ponce. Automatic annotation of human actions in video. In *ICCV*, 2009. 2, 3

[10] Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992. 6

[11] Paul Ekman. Basic emotions. *Handbook of cognition and emotion*, 98(45-60):16, 1999. 2

[12] Jiarui Gao, Yanwei Fu, Yu-Gang Jiang, and Xiangyang Xue. Frame-transformer emotion classification network. In *ICMR*, 2017. 1, 8

[13] Mingfei Gao, Yingbo Zhou, Ran Xu, Richard Socher, and Caiming Xiong. Woad: Weakly supervised online action detection in untrimmed videos. In *CVPR*, 2021. 2

[14] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *ICML*, 2006. 3

[15] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *CVPR*, 2018. 3

[16] De-An Huang, Li Fei-Fei, and Juan Carlos Niebles. Connectionist temporal modeling for weakly supervised action labeling. In *ECCV*, 2016. 3

[17] Linjiang Huang, Liang Wang, and Hongsheng Li. Foreground-action consistency network for weakly supervised temporal action localization. In *ICCV*, 2021. 2

[18] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *TPAMI*, 35(1):221–231, 2012. 6

[19] Guoli Jia and Jufeng Yang. S 2-ver: Semi-supervised visual emotion recognition. In *ECCV*, 2022. 2

[20] Yu-Gang Jiang, Baohan Xu, and Xiangyang Xue. Predicting emotions in user-generated videos. In *AAAI*, 2014. 1, 2, 6

[21] Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008. 3

[22] Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. Context-aware emotion recognition networks. In *ICCV*, 2019. 6

[23] Li-Jia Li, Hao Su, Li Fei-Fei, and Eric P Xing. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *NIPS*, 2010. 2

[24] Zheng Lian, Bin Liu, and Jianhua Tao. Smin: Semi-supervised multi-modal interaction network for conversational emotion recognition. *TAC*, 2022. 2

[25] Wang Luo, Tianzhu Zhang, Wenfei Yang, Jingen Liu, Tao Mei, Feng Wu, and Yongdong Zhang. Action unit memory network for weakly supervised temporal action localization. In *CVPR*, 2021. 2

[26] Junwei Ma, Satya Krishna Gorti, Maksims Volkovs, and Guangwei Yu. Weakly supervised action selection learning in video. In *CVPR*, 2021. 2

[27] Sijie Mai, Ying Zeng, Shuangjia Zheng, and Haifeng Hu. Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. *TAC*, 2022. 2

[28] Navonil Majumder, Devamanyu Hazarika, Alexander Gelbukh, Erik Cambria, and Soujanya Poria. Multimodal sentiment analysis using hierarchical fusion with context modeling. *KBS*, 161:124–133, 2018. 2

[29] Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. In *AAAI*, 2020. 2

[30] Phuc Nguyen, Ting Liu, Gautam Prasad, and Bohyung Han. Weakly supervised action localization by sparse temporal pooling network. In *CVPR*, 2018. 3

[31] John Edison Arevalo Ovalle, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. Gated multimodal units for information fusion. In *ICLR (Workshop)*, 2017. 6

[32] Lei Pang and Chong-Wah Ngo. Mutlimodal learning with deep boltzmann machine for emotion prediction in user generated videos. In *ICMR*, 2015. 1

[33] Lei Pang, Shiai Zhu, and Chong-Wah Ngo. Deep multimodal learning for affective analysis and retrieval. *TMM*, 17(11):2008–2020, 2015. 6

[34] Sujoy Paul, Sourya Roy, and Amit K Roy-Chowdhury. W-talc: Weakly-supervised temporal activity localization and classification. In *ECCV*, 2018. 3, 7, 8

[35] Robert Plutchik. A general psychoevolutionary theory of emotion. In *Theories of Emotion*, pages 3–33. Elsevier, 1980. 6

[36] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. Context-dependent sentiment analysis in user-generated videos. In *ACL*, 2017. 1

[37] Haonan Qiu, Liang He, and Feng Wang. Dual focus attention network for video emotion recognition. In *ICME*, 2020. 6

[38] Harold Schlosberg. Three dimensions of emotion. *Psychological Review*, 61(2):81, 1954. 2

[39] Kah Phooi Seng and Li-Minn Ang. Video analytics for customer emotion and satisfaction at contact centers. *IEEE Transactions on Human-Machine Systems*, 48(3):266–278,

2018. 1

[40] Dongyu She, Jufeng Yang, Ming-Ming Cheng, Yu-Kun Lai, Paul L Rosin, and Liang Wang. Wscnet: Weakly supervised coupled networks for visual sentiment classification and detection. *TMM*, 22(5):1358–1371, 2019. 2

[41] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Adasgn: Adapting joint number and model size for efficient skeleton-based action recognition. In *ICCV*, 2021. 2

[42] Zheng Shou, Hang Gao, Lei Zhang, Kazuyuki Miyazawa, and Shih-Fu Chang. Autoloc: Weakly-supervised temporal action localization in untrimmed videos. In *ECCV*, 2018. 7, 8

[43] Karan Sikka, Karmen Dykstra, Suchitra Sathyanarayana, Gwen Littlewort, and Marian Bartlett. Multiple kernel learning for emotion recognition in the wild. In *ICMI*, 2013. 2

[44] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *ICCV*, 2017. 3

[45] Chen Sun, Sanketh Shetty, Rahul Sukthankar, and Ram Nevatia. Temporal localization of fine-grained actions in videos by domain transfer from web images. In *ACM MM*, 2015. 3

[46] Teng Sun, Wenjie Wang, Liqaing Jing, Yiran Cui, Xuemeng Song, and Liqiang Nie. Counterfactual reasoning for out-of-distribution multimodal sentiment analysis. In *ACM MM*, 2022. 2

[47] Guoyun Tu, Yanwei Fu, Boyang Li, Jiarui Gao, Yu-Gang Jiang, and Xiangyang Xue. A multi-task neural approach for emotion attribution, classification and summarization. *TMM*, 22(1):148–159, 2020. 5

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017. 4

[49] Lijuan Wang, Guoli Jia, Ning Jiang, Haiying Wu, and Jufeng Yang. Ease: Robust facial expression recognition via emotion ambiguity-sensitive cooperative networks. In *ACM MM*, 2022. 1

[50] Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. Untrimmednets for weakly supervised action recognition and detection. In *CVPR*, 2017. 3, 7, 8

[51] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018. 4, 7

[52] Jie Wei, Xinyu Yang, and Yizhuo Dong. User-generated video emotion recognition based on key frames. *Multimedia Tools and Applications*, pages 1–19, 2021. 1

[53] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *CVPR*, 2017. 5

[54] Baohan Xu, Yanwei Fu, Yu-Gang Jiang, Boyang Li, and Leonid Sigal. Heterogeneous knowledge transfer in video emotion recognition, attribution and summarization. *TAC*, 9(2):255–270, 2016. 1, 2, 6

[55] Baohan Xu, Yingbin Zheng, Hao Ye, Caili Wu, Heng Wang, and Gufei Sun. Video emotion recognition with concept selection. In *ICME*, 2019. 6

[56] Jufeng Yang, Dongyu She, Yu-Kun Lai, Paul L Rosin, and Ming-Hsuan Yang. Weakly supervised coupled networks for visual sentiment analysis. In *CVPR*, 2018. 2

[57] Jufeng Yang, Dongyu She, and Ming Sun. Joint image emotion classification and distribution learning via deep convolutional neural network. In *IJCAI*, 2017. 1

[58] Xumin Yu, Yongming Rao, Wenliang Zhao, Jiwen Lu, and Jie Zhou. Group-aware contrastive regression for action quality assessment. In *ICCV*, 2021. 2

[59] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. *EMNLP*, 2017. 6

[60] Can Zhang, Meng Cao, Dongming Yang, Jie Chen, and Yuexian Zou. Cola: Weakly-supervised temporal action localization with snippet contrastive learning. In *CVPR*, 2021. 2

[61] Haimin Zhang and Min Xu. Recognition of emotions in user-generated videos with kernelized features. *TMM*, 20(10):2824–2835, 2018. 2, 6

[62] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas S Huang. Adversarial complementary learning for weakly supervised object localization. In *CVPR*, 2018. 5

[63] Zhicheng Zhang and Jufeng Yang. Temporal sentiment localization: Listen and look in untrimmed videos. In *ACM MM*, 2022. 1

[64] Sicheng Zhao, Guoli Jia, Jufeng Yang, Guiguang Ding, and Kurt Keutzer. Emotion recognition from multiple modalities: Fundamentals and methodologies. *SPM*, 38(6):59–73, 2021. 2

[65] Sicheng Zhao, Yunsheng Ma, Yang Gu, Jufeng Yang, Tengfei Xing, Pengfei Xu, Runbo Hu, Hua Chai, and Kurt Keutzer. An end-to-end visual-audio attention network for emotion recognition in user-generated videos. In *AAAI*, 2020. 1, 2, 3, 5, 6, 7, 8

[66] Sicheng Zhao, Xingxu Yao, Jufeng Yang, Guoli Jia, Guiguang Ding, Tat-Seng Chua, Bjoern W Schuller, and Kurt Keutzer. Affective image content analysis: Two decades review and new perspectives. *TPAMI*, 44(10):6729–6751, 2021. 1

[67] Jia-Xing Zhong, Nannan Li, Weijie Kong, Tong Zhang, Thomas H. Li, and Ge Li. Step-by-step erasion, one-by-one collection: A weakly supervised temporal action detector. In *ACM MM*, 2018. 7, 8