Representation Learning through Multimodal Attention and Time-Sync Comments for Affective Video Content Analysis

Jicai Pan
panjc@mail.ustc.edu.cn
University of Science and Technology of China
Hefei, Anhui, China

Shangfei Wang*
sfwang@ustc.edu.cn
University of Science and Technology of China
Hefei, Anhui, China

Lin Fang
clivefang@ustc.edu
University of Science and Technology of China
Hefei, Anhui, China

ABSTRACT

Although temporal patterns inherent in visual and audio signals are crucial for affective video content analysis, they have not been thoroughly explored yet. In this paper, we propose a novel Temporal-Aware Multimodal (TAM) method to fully capture the temporal information. Specifically, we design a cross-temporal multimodal fusion module that applies attention-based fusion to different modalities within and across video segments. As a result, it fully captures the temporal relations between different modalities. Furthermore, a single emotion label lacks supervision for learning representation of each segment, making temporal pattern mining difficult. We leverage time-synchronized comments (TSCs) as auxiliary supervision, since these comments are easily accessible and contain rich emotional cues. Two TSC-based self-supervised tasks are designed: the first aims to predict the emotional words in a TSC from video representation and TSC contextual semantics, and the second predicts the segment in which the TSC appears by calculating the correlation between video representation and TSC embedding. These self-supervised tasks are used to pre-train the cross-temporal multimodal fusion module on a large-scale video-TSC dataset, which is crawled from the web without labeling costs. These self-supervised pre-training tasks prompt the fusion module to perform representation learning on segments including TSC, thus capturing more temporal affective patterns. Experimental results on three benchmark datasets show that the proposed fusion module achieves state-of-the-art results in affective video content analysis. Ablation studies verify that after TSC-based pre-training, the fusion module learns more segments' affective patterns and achieves better performance.

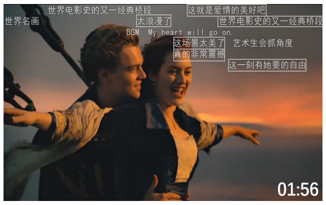
CCS CONCEPTS

• **Computing methodologies → Hierarchical representations**.

KEYWORDS

affective computing, video content analysis, multimodal fusion, vision and language

ACM Reference Format:
Jicai Pan, Shangfei Wang, and Lin Fang. 2022. Representation Learning through Multimodal Attention and Time-Sync Comments for Affective Video Content Analysis. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3503161.3548018



TSC List in a TITANIC Clip
(In English)

01:51	*So romantic*
01:52	*This scene is so beautiful*
01:52	*Really impressive*
01:53	*This is the beauty of love*
01:54	*Classic clip in world film history*
01:55	*She is freedom at this moment*

Figure 1: The left is a Titanic clip at 1min56s; the right shows the time-sync comments appearing between 1min51s and 1min55s. Several viewers commented on feelings such as romance, awe, love, and freedom.

1 INTRODUCTION

Affective video content analysis aims to predict the emotions that viewers are expected to be evoked or really evoked when watching videos. The development of video-sharing websites has led to the proliferation of videos on social websites such as Youtube and Bilibili. Managing this vast number of videos is a challenge. Classifying videos by their induced emotion is a good solution with multiple benefits. First, video retrieval through emotional keywords is easy and accurate. Second, website managers can also utilize the induced emotions of the videos to enhance the recommendation system and improve viewers' experience. Third, understanding affective video content can help video generators make more appealing videos. Last, affective video content analysis could also be used to detect extreme emotional videos, allowing government regulators to take preventive measures quickly. Therefore, affective video content analysis has attracted attention over the years [24, 39].

While computer vision methods have become more advanced, affective video content analysis still faces several challenges. First, the visual and audio information need to be fused efficiently, since both can provoke the viewer's emotions. Existing methods mainly adopt decision-level or feature-level fusion to integrate visual and audio signals. The former directly combines analysis results from visual and audio features, ignoring dependencies across modalities. The latter typically captures the temporal features within each modality and fuses them into a joint feature. However, since most

*Corresponding Author.

MM '22, October 10–14, 2022, Lisboa, Portugal
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3548018

feature-level fusion methods capture temporal and cross-modal dependencies separately, they don't fully capture the temporal dependencies between visual and audio signals. This paper introduces a cross-temporal multimodal fusion module to solve this issue. The proposed fusion module applies the self-attention operation to different modalities within each video segment as well as across segments. Consequently, all temporal dependencies across visual and audio modalities are exploited for affective video content analysis.

Second, current one-label-per-video supervision is insufficient, since affective states vary within different segments of the same video. For instance, a surprise video could end with a happy atmosphere setting. This kind of affective transition can be used as a temporal pattern for affective content understanding. However, the supervision provided by a single emotional label is too limited to describe the affective states of all video segments, making it difficult to mine temporal patterns. We address this by introducing time-sync comments (TSCs) as auxiliary supervision to pre-train the cross-temporal multimodal fusion module. Time-sync comments are brief timestamped viewer comments containing emotional feelings. As shown in Fig. 1, the TSCs in this famous movie clip express multiple emotions aroused by this scene: romance, awe, love, and freedom. TSCs can provide emotional and temporal cues for affective video content analysis. We design two TSC-based self-supervised tasks: emotional word predicting and appearing time predicting. The former uses the video representation extracted by the fusion module and TSC contextual semantics to predict the emotional words in TSC, and the latter predicts which TSC goes with which segment by calculating the similarities between video representation and TSC embedding. These two tasks make full use of TSC semantics and temporal cues to enhance segment-level representation learning. Videos with intensive TSCs are easily accessible on the Internet. The proposed fusion module is pre-trained on a large-scale video-TSC dataset collected from the web, requiring no manual annotation. The TSCs are discarded during affective video analysis inference.

The temporal-aware multimodal method is evaluated on three popular benchmark datasets: VE-8 [13], YF-6 [27], and LIRIS-ACCEDE [2]. Experimental results demonstrate that the fusion module achieves state-of-the-art results, extracting more discriminative representation after TSC-based pre-training.

In summary, the contributions of the proposed temporal-aware multimodal method are as follows: First, we design a cross-temporal multimodal fusion module to learn the temporal dependencies between visual and audio signals for affective video content analysis. Second, we propose two TSC-based self-supervised pre-training tasks to enhance segment-level representation learning. Third, we conduct extensive experiments on three affective video datasets to demonstrate the effectiveness of the proposed method.

2 RELATED WORK

2.1 Affective Video Content Analysis

Multimodal fusion of affective video content analysis can be divided into two types: decision-level fusion and feature-level fusion. Decision-level fusion combines the results from different classifiers, ignoring the dependencies between visual and audio features. For example, Acar et al. [1] learned visual and audio features separately and then employed three multi-class support vector machines to obtain affective predictions. Feature-level fusion combines visual and audio features and feeds them jointly to a classifier or regressor. For example, Xu et al. [28] selected each modal feature based on emotional concepts, and summed these selected features into a joint feature for emotion classification. Qiu et al. [20] simply concatenated action and scene features as a whole and then input them into a dual attention network to focus on emotion-related frames. Wei et al. [25] also concatenated object and scene features, with a focus on estimating the affective saliency value of frames. Zhao et al. [38] integrated spatial, channel-wise, and temporal attention into a visual extractor, and temporal attention into an audio extractor, then concatenated the output visual and audio features into a joint emotional feature. However, these methods used a simple fusion strategy, summing or concatenating the visual and audio features while ignoring inherent interactions among them.

Some feature-level fusion methods use complex strategies to mine the dependencies between visual and audio signals. For instance, Gan et al. [9] used a regression Bayesian network to capture the high-order dependencies between low-level visual and audio features, ignoring temporal patterns. Qi et al. [19] used an attention mechanism to aggregate the temporal features in each modality, then aligned the visual and audio features by jointly mapping them into a common space. Mittal et al. [18] used a long short-term memory (LSTM) encoder to learn the temporal features and used a co-attention mechanism to calculate correlation scores between the pairwise modalities. They then weighted and summed these multimodal temporal features using correlation scores. Gao et al. [10] proposed a synchronous modal-temporal attention block to capture the visual and audio relations within each moment, then used LSTM to learn the temporal dependencies within each modality. However, these methods either discard temporal relations or learn temporal and cross-modal dependencies separately, thereby ignoring inherent dependencies between temporal elements of visual and audio signals, which are essential for effective multimodal fusion in videos [16].

To mine these dependencies, we design a cross-temporal multimodal fusion module that employs self-attention to learn the pairwise modalities' relations within each video segment and across different segments. This module simultaneously learns relations between all the video segments of different modalities, so it can fully capture the temporal dependencies between visual and audio signals.

2.2 Video Analysis with TSCs

TSCs have great research significance for video understanding. It can be used for video tagging [17, 26, 33], video description [30], and video recommendation [5, 32], among other things. TSCs are viewers' real-time emotional expressions to video content. Their semantics and temporal cues indicate the induced emotion of the video and the moment of the emotion burst. Hence, TSC has great potential for affective video content analysis. However, this potential has not been successfully explored. To the best of our knowledge, only Li et al. [14] leveraged TSC for video emotion recognition. They first utilized canonical correlation analysis to maximize the mutual information between visual and TSC textual features, and then used LSTM to separately capture temporal dependencies within

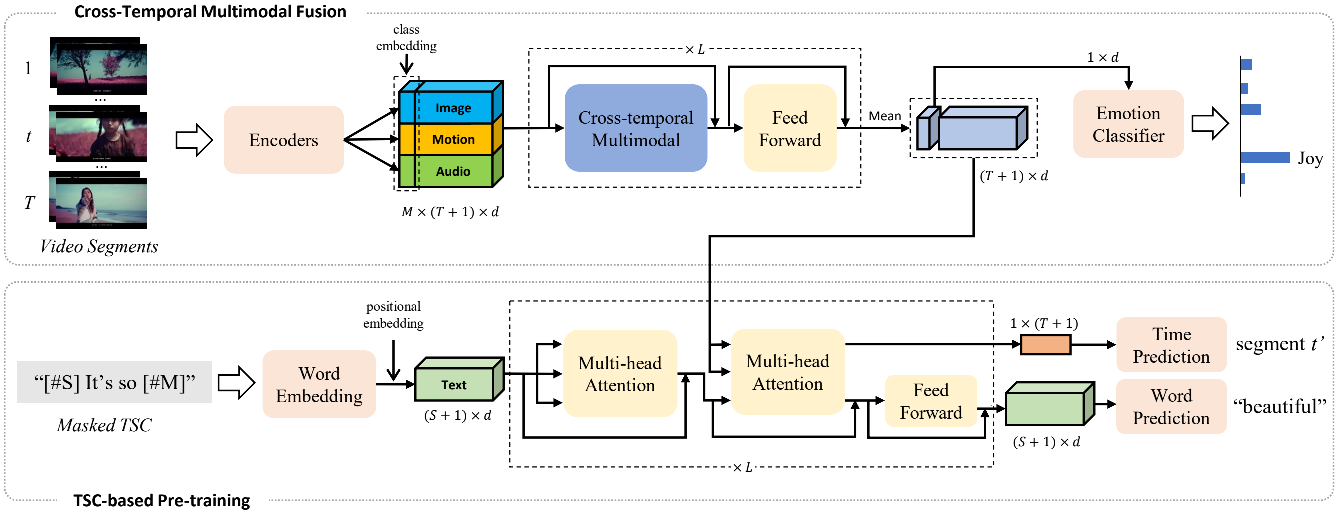


Figure 2: The overall structure of the temporal-aware multimodal method. Our method can be divided into two parts: cross-temporal multimodal fusion and TSC-based pre-training. Image, motion, and audio features are extracted from three video encoders. The input of the pre-training is a masked TSC, where $[\#S]$ and $[\#M]$ are the start token and mask token, respectively. The outputs of the pre-training tasks represent that the TSC is possible to appear at segment t' and the masked word is "beautiful".

visual and TSC modalities. The output features of visual and TSC are concatenated as a joint feature for video emotion recognition. However, their method requires emotion-labeled video-TSC data during training, and the expensive labeling costs limit the number of training samples. Moreover, this method needs to simultaneously input video and TSC in the inference stage, limiting its application in video analysis.

We design two self-supervised pre-training tasks to mine the semantic and temporal cues from TSCs to enhance segment-level representation learning. These tasks are designed to force the cross-temporal multimodal fusion module to learn the representation of segments in which TSCs appear. Specifically, the emotional word predicting task uses video representation and TSC contextual semantics to fill the masked TSC with emotional words. The appearing time predicting task computes the similarities between video representation and TSC embedding to predict the segment in which the TSC appears. We collect a large-scale video-TSC dataset to pre-train the cross-temporal multimodal fusion module without manual annotation. We also fine-tune the pre-trained fusion module on affective video datasets for affective video analysis. TSCs are not needed in the fine-tuning and inference stages.

3 PROBLEM STATEMENT

Suppose we have an affective video dataset $\mathcal{D}_a = \{\mathcal{V}_i^a, y_i\}_{i=1}^{N_v}$, and a video-TSC dataset $\mathcal{D}_c = \{\mathcal{V}_i^c, \mathcal{T}_i\}_{i=1}^{N_c}$. The $\mathcal{D}_a$ contains N_v videos with affective labels y. The $\mathcal{D}_c$ contains N_c videos with TSC sets. $\mathcal{T}_i = \{t_{ij}, [w_{ij}^1, w_{ij}^2, ..., w_{ij}^S]\}_{j=1}^{N_i}$ is the TSC set of the i-th video. t_{ij} is the appearing segment of the j-th TSC. S is the maximum number of words in the TSCs. N_i is the number of TSCs in $\mathcal{V}_i^c$. Our goal is to pre-train a network in a self-supervised manner on the video-TSC dataset $\mathcal{D}_c$, then fine-tune this network with the affective labels on the affective video dataset $\mathcal{D}_a$. Only videos are inputted into the network to predict the affective labels during inference.

4 METHODOLOGY

In this section, we introduce the proposed temporal-aware multimodal (TAM) method in detail. As shown in Figure 2, the proposed TAM method consists of two parts: cross-temporal multimodal fusion and TSC-based pre-training. The former helps capture the temporal relations between different modalities, while the latter designs two self-supervised tasks forcing the backbone to fully capture the affective pattern throughout video segments.

As shown in the upper part of Fig. 2, each video is divided into T segments for affective video content analysis. For each video segment, we use $M = 3$ video encoders to extract the image, motion, and audio features $X \in \mathbb{R}^{M \times T \times d}$, where d represents the feature dimension. A cross-temporal multimodal fusion module is designed to merge the feature X into video representation f, then inputs it into a classifier or regressor to predict the affective label y.

4.1 Cross-Temporal Multimodal Fusion

For efficient affective video content analysis, the fusion module needs to fully capture the correlation between the features of different modalities and temporal segments. Recently, multi-head self-attention operation in Transformer [22] achieves outstanding performance in various vision tasks [8] by learning long-range dependencies between sequences. Inspired by this, we adopt a variant of the self-attention operation to learn both the temporal and modal dependencies to fully explore the video information.

Similar to ViT [8], we first prepend an extra learnable embedding $X_0^m \in \mathbb{R}^{1 \times d}$ to the m-th modal sequence for learning the comprehensive video feature, and add a position embedding $P_e \in \mathbb{R}^{(T+1) \times d}$

to each modal sequence to retain temporal information as follows:

$$Z^m = [X_0^m; X_1^m; X_2^m; ...; X_T^m] + P_e \tag{1}$$

where $m \in \{1, 2, ..., M\}$. The feature Z is input into a cross-temporal multimodal (CTM) fusion module for multimodal fusion. The fusion module consists of an inner-modal attention and a cross-modal attention, as shown in Fig. 3. Specifically, the feature of each modality are mapped into query, key, and value domains:

$$Q^m, K^m, V^m = Z^m W_q, Z^m W_k, Z^m W_v \tag{2}$$

where $W_q, W_k, W_v \in \mathbb{R}^{d\times d}$ are learnable parameter matrices for the m-th modality and Q^m, K^m, and V^m are the query, key, and value matrices respectively.

The inner-modal attention encodes the temporal correlations between the features within the same modality by matching their query and key matrices. The inner-modal fusion results are calculated as:

$$I^m = \sigma(\frac{Q^m K^{m\top}}{\sqrt{d}})V^m \tag{3}$$

where $K^{m\top}$ is the transposed key matrix and $I^m \in \mathbb{R}^{(T+1)\times d}$ is the inner-modal fusion result of the m-th modal features, $\sigma(\cdot)$ represents the softmax function.

Cross-modal attention aims to capture the temporal correlation between pairwise features of different modalities. We match the query matrix Q^m of the m-th modality with the key matrices of all other modalities to learn the correlation weights, and then average the weighted features from different modalities to get cross-modal fusion results, as follows:

$$C^m = \frac{1}{M-1}\sum_{n\neq m}\sigma(\frac{Q^m K^{n\top}}{\sqrt{d}})V^n \tag{4}$$

where $n \in \{1, ..., M\}$, $C^m \in \mathbb{R}^{(T+1)\times d}$ is the cross-modal fusion result of the m-th modal features.

The cross-temporal multimodal fused features are obtained by concatenating the inner-modal attention and cross-modal attention as follows:

$$O^m = (I^m \oplus C^m)W_o \tag{5}$$

where $\oplus$ is matrix concatenation along the last dimension, $W_o \in \mathbb{R}^{2d\times d}$ is a learnable parameter matrix.

We fully capture the long-range dependencies between the features of different modalities and temporal segments by stacking the proposed cross-temporal multimodal fusion module L times, along with L feed-forward network (FFN) [22]. Finally, we average the output of the last cross-temporal multimodal fusion module to obtain the final video representation:

$$f = \frac{1}{M}\sum_{m=1}^{M} O^m \tag{6}$$

During TSC-based pre-training, the whole video representation is input into the video-TSC decoder and optimized by two self-supervised tasks. During video affective analysis, we input the first temporal video representation $f^0 \in \mathbb{R}^{1\times d}$ into a classifier or regressor to predict the affective result for the video.

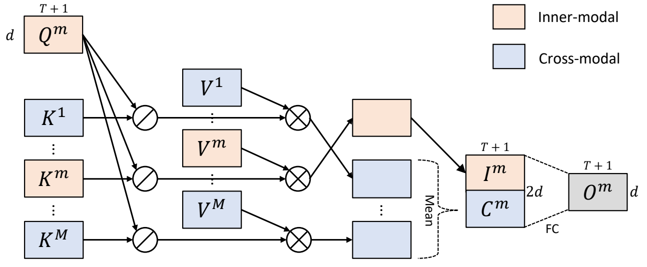


Figure 3: The cross-temporal multimodal fusion module consists of cross-modal and inner-modal attentions. $\oslash(Q, K) = \sigma(QK^\top/\sqrt{d})$, $\otimes$ is matrix multiplication, and FC represents the fully-connected layer.

4.2 TSC-Based Pre-Training

We predict the emotion label y_i' of the video by fusing the features of different modalities and temporal segments. The cross-temporal multimodal fusion module is supervised by the target emotion label y_i during training. However, the supervision provided by a single emotional label is insufficient to describe the affective states throughout video segments, making temporal pattern mining difficult. To address this problem, we design two TSC-based self-supervised tasks to pre-train the fusion module for improved mining of temporal affective patterns.

The procedure of the TSC-based self-supervised pre-training is shown in the lower part of Fig. 2. Specifically, the video-TSC decoder is a Transformer decoder structure, which inputs the word embeddings of TSC and video representation and outputs the TSC semantic feature with the video context. The emotional word predicting task aims to predict the actual emotional word in a TSC within the video and TSC context. It enhances the video feature extractor to learn emotional features. The appearing time predicting task aims to align the video representation and TSC semantics in temporal. It forces the video feature extractor to focus on the segment where the TSC appears.

The masked language model (MLM) [7] is highly successful at learning sentence semantics. Inspired by this, we design the emotional word predicting task based on MLM. This task uses the guidance of the video features and the features from other non-masked words to predict the masked word in each TSC. Unlike the random masking strategy of the original MLM, we mask emotional words in each TSC for emotional content learning. We first use a Chinese emotion lexicon [29] to find the emotional words in a TSC, then use a special word [#M] to mask these emotional words. This masked TSC is input into the video-TSC decoder accompanied by video representation from the cross-temporal multimodal fusion module. The output embedding of the [#M] word is input into a classifier to predict the original word. The loss of emotional word predicting task can be formulated as:

$$\mathcal{L}_{wp} = -\frac{1}{|\mathcal{T}|}\sum_{i=1}^{N_c}\sum_{j=1}^{N_i}\log P(w_{ij}^m | w_{ij}^{\backslash m}, f_i) \tag{7}$$

where $|\mathcal{T}|$ is the total TSC number of the video-TSC dataset, m indicates that the m-th word is the masked emotional word, $\backslash m$

means the other non-masked words, and f_i represents the video representation of the i-th video.

To enhance the temporal level video representation learning, we align the TSC embedding and the video segment feature through the appearing time predicting task. This task aims to predict the appearing temporal segment of the masked TSC. Specifically, we use $g \in \mathbb{R}^{1\times d}$ to represent the output embedding of the first word from the first multi-head attention block. In the second multi-head attention block, we calculate the appearing segment prediction as follows:

$$P(t|g,f) = \frac{\exp(\hat{g} \cdot \tilde{f^t}^{\mathsf{T}})}{\sum_k \exp(\hat{g} \cdot \tilde{f^k}^{\mathsf{T}})} \tag{8}$$

where $k \in \{1, 2, ..., T\}$, $\hat{g}$ and $\tilde{f}$ are the projections of g in query domain and f in key domain, respectively. Therefore, the loss of appearing time predicting task can be calculated as:

$$\mathcal{L}_{tp} = -\frac{1}{|\mathcal{T}|} \sum_{i=1}^{N_c} \sum_{j=1}^{N_i} \log P(t_{ij}|g_{ij}, f_i) \tag{9}$$

The emotional word predicting task utilizes the abundant emotional cues in TSC as auxiliary supervision for multimodal representation learning, while the appearing time predicting task forces this learning to focus on the appearing temporal segment of TSC. Consequently, the video encoders and fusion modules learn to fully capture affective information throughout the temporal segments.

The total loss of the TSC-based pre-training is formulated as:

$$\mathcal{L} = \lambda_w \mathcal{L}_{wp} + \lambda_t \mathcal{L}_{tp} \tag{10}$$

where λ_w, λ_t are two trade-off parameters that weigh the importance of these two tasks.

After TSC-based pre-training, the TSC-decoder is removed for affective video content analysis without TSC input. Finally, we fine-tune the pre-trained cross-temporal multimodal module and the classifier/regressor on the affective video dataset.

5 EXPERIMENTS

This section first introduces the three public benchmark datasets and the self-collected video-TSC dataset. Then we perform several ablation studies to verify the effectiveness of our design. Lastly, we compare the proposed method to other state-of-the-art methods.

5.1 Datasets

The VideoEmotion-8 (VE-8) [13] dataset is collected from YouTube and Flickr, and it contains a total of 1,101 videos with an average duration of 107 seconds. These videos are labeled with one of eight emotions: *anger, anticipation, disgust, fear, joy, sadness, surprise, or trust.* Each category contains a minimum of 100 videos. We follow the common experimental setup [13, 38], randomly splitting the dataset for ten runs with 2/3 training and 1/3 testing, and report the average results of the ten runs.

The YouTube/Flickr-EkmanSix (YF-6) [27] dataset consists of 1,637 videos collected from YouTube and Flickr. The average duration is 112 seconds. The videos are labeled with one of six basic emotion categories: *anger, disgust, fear, joy, sadness, or surprise.* There are at least 221 videos in each category. We use the public splitting of 819 videos for training and 818 for testing.

The LIRIS-ACCEDE [2] dataset is the largest dataset for video affective analysis, consisting of 9,800 videos extracted from 160 movies. The videos last between 8 and 12 seconds. There are two tasks based on this dataset. In the MediaEval2015 affective impact of movies task, a total of 10,900 (1,100 additional) videos are used for classification. These videos are split into 6,144 videos for training and 4,756 videos for testing. For each video, the ground truth consists of the arousal class (calm-neutral-active) and the valence class (negative-neutral-positive). MediaEval2016 emotional impact of movies is a regression task and includes 11,000 videos (1,200 additional) split into 9,800 training videos and 1,200 testing videos. Each video has the absolute affective scores of valence and arousal. Consistent with prior methods [9, 35], we use prediction accuracy (ACC) for MediaEval2015 evaluation, use Mean Square Error (MSE) and Pearson Correlation Coefficient (PCC) for MediaEval2016 evaluation.

The Video-TSC dataset is collected from the Chinese video website Bilibili.[1] We crawled 7,000 videos containing intensive TSCs from the life, short film, and popularity sections published between January 2018 and December 2021. Most of the TSCs are brief Chinese sentences with around 10 words. We divide these videos into segments to facilitate analysis. Specifically, referring to [14], we use the K-means algorithm to cluster the appearing time of the TSC with each cluster corresponding to the video segment to be split. For each video, we set K as the video duration divided by 30 so that each video segment is around 30 seconds. Furthermore, to balance the sentiment in these videos, we predict the sentiment score (ranging from 0 to 1, negative to positive) of each TSC and then take the average of all the TSC sentiment scores in the video segment as the final sentiment score. These sentiment scores are only used to pick out an equal number of positive and negative videos, and are not used in the training stages. We selected the 8,000 most negative and 8,000 most positive video segments, with a total average duration of 27.6 seconds. There are a total of 6,831,524 TSCs in these video segments.

5.2 Implementation Details

Video features are extracted using three video encoders: CLIP-enhanced ViT [8, 21] for image features, ResNet3D [11] for motion features, and VGGish [12] for audio features. The three models are pre-trained on large-scale datasets via image text aligning, action recognition, and audio classification tasks. Specifically, we choose Base Vision Transformer with 32 layers (ViT-B/32) pre-trained by CLIP as the image encoder. The input image is 224 × 224 pixels, and the output is a 512-dimensional feature. The ResNet3d needs 16 consecutive 112 × 112 frames as the input and outputs a 2048-dimensional feature. The audio is first converted to the Mel-frequency cepstral coefficient (MFCC), then input into the VGGish model to obtain a 128-dimensional feature. These outputs are input into fully-connected layers to be separately mapped to the same 768-dimensional space. The parameters of these three encoders do not participate in gradient optimization.

[1] https://www.bilibili.com

For each video, we randomly sample T consecutive video segments, each containing 16 frames. All frames and audio MFCCs corresponding to the T segments are directly input into the motion and audio encoders. In each segment, one frame is selected and input into the image encoder. The number of cross-temporal multimodal fusion layers is $L = 12$. Models are optimized using an Adam optimizer with a learning rate of 0.0001 and weight decay of 0.005. The models are trained with batch size 8 for 100 epochs and $T = 64$ for classification tasks on VE-8 and YF-6. For MediaEval2015 and MediaEval2016 tasks, the batch size is 16 and the training epoch is 20, and $T = 16$. For each task, we randomly split 15% of the training set off as a validation set to choose the best hyperparameters.

To utilize the emotional and content-related TSCs for better video-TSC pre-training, we preprocess TSCs as follows: First, we delete the TSCs that do not contain emotional words. The emotional words are provided by the Chinese emotion dictionary [29], which contains 27,466 emotional words. The TSC with emotional words significantly supports emotion classification and recognition [31]. Second, too short (less than 3 words) or too long (more than 15 words) TSCs are deleted. Usually, too short TSCs often fail to convey the actual emotions, and too long TSCs are generally disorganized words or irrelevant to the current video content [17]. Third, TSCs containing special characters and numbers are filtered out. Due to the lack of special character sentiment corpus, it is difficult to parse out the sentiment content of these characters. Although these preprocessing cannot ensure all TSCs are of high quality, we can mine the major emotional information from these TSCs according to the TSC herding effect [32], namely, most of TSCs are highly semantic relevant and time-related. Finally, removing the duplicate TSCs that appear within a second yields 891,400 TSCs for video-TSC pre-training. Each TSC is tokenized by the Chinese tool Jieba[2]. The number of TSC decoder layers is $L = 8$. In the pre-training phase, we randomly divide the video-TSC dataset into training and testing sets at a ratio of 4:1. An Adam optimizer is used and the total training epoch is 200; the initial learning rate is 0.001, and it decays by 0.1 every 50 epochs. During fine-tuning, the learning rates are 0.00001 and 0.0001 for the pre-trained cross-temporal multimodal fusion module and the classifier (or regressor), respectively. The fine-tuning epoch is 50.

5.3 Baseline

We compare our cross-temporal multimodal fusion strategy to three other fusion strategies drawn from [3, 15, 38]. The detailed implementations are as follows:

Simple concatenation [38] is stacked with 12 identical layers. Each layer consists of an eight-head self-attention, a feed-forward network, and the residual connection around every two sub-layers. The image, motion, and audio features are separately input into three parallel fusion modules. Then the first outputs of these three modules are concatenated as the fused feature. This simple concatenation strategy uses the attention mechanism to focus on temporal relations within each modality, but disregards the interactions between different modalities.

Divide attention [3] is formed by inserting an eight-head self-attention between the two sub-layers of the simple concatenation fusion module. First, the temporal features of each modality are input into the first self-attention. Then, the multimodal features of each segment are input into the second self-attention. Thus the temporal relations and cross-modal dependencies are separately captured.

Joint attention [15] is the same as the simple concatenation fusion module. First, all temporal features of the three modalities are concatenated as a joint sequence. Then the joint sequence is input into the fusion module for full pairwise attention between segments and modalities. This fusion strategy mixes up the temporal and cross-modal relations.

5.4 Ablation Study

We perform an ablation study featuring different combinations of the modalities to show the necessity of each modality. Table 1 shows the video emotion recognition accuracy of different combinations of modalities. From the table, we can have the following observations. First, combining three modalities achieves the best performance of 56.04% and 60.64% on VE-8 and YF-6. Second, the smallest gap between all three modalities combined and the other combinations is 3.02% and 1.47%. These results verify that all three modalities contribute to the performance of the model. Thus it is essential for affective video content analysis to design an effective multimodal fusion module.

To validate the effectiveness of the proposed cross-temporal multimodal (CTM) fusion module, we perform an ablation study on different multimodal fusion designs. Specifically, the fusion strategies mentioned in Sec. 5.3 are set as the baselines and compared to the CTM module on VE-8 and YF-6 datasets. Results are shown in the upper part of Table 2. From the table, we can make two major observations. First, our CTM fusion module achieves the highest emotion classification accuracy on both VE-8 and YF-6. It outperforms the simple concatenation and divide attention fusion modules by a large margin. This is because these modules ignore the temporal dependencies between different modalities. Secondly, by capturing all relations between video segments and modalities, joint attention fusion achieves admirable performance as well. However, our CTM fusion module still has a 1.92% and 0.36% advantage over it. This is likely because the joint attention fusion employs the full pairwise attention on all temporal and modal features, and the redundancy of visual and audio information weakens its capability.

Ablation experiments are carried out on the VE-8 and YF-6 datasets to analyze the contributions of different attentions of our fusion module. The lower part of Table 2 shows the results of inner-modal attention only, cross-modal attention only, and the combination of these attentions. The inner-modal attention strategy aims to mine the temporal dependencies within each modality, and the cross-modal one aims to capture the temporal relations across different modalities. Combining these two attentions yields the highest accuracies of 56.04% and 60.64% on VE-8 and YF-6, which is 1.64% and 2.2% higher than the inner-modal attention, as well as 9.89% and 12.35% higher than the cross-modal attention. Although the cross-modal attention is not competitive compared to the inner-modal attention, it complements inner-modal dependencies. This verifies that both inner-modal and cross-modal attentions

[2] https://github.com/fxsjy/jieba

Table 1: Video emotion recognition accuracy (%) with different modalities

Image	Motion	Audio	VE-8	YF-6
✓			52.75	57.82
	✓		46.15	48.04
		✓	40.38	40.59
✓	✓		50.82	57.82
✓		✓	53.02	59.17
	✓	✓	45.05	52.69
✓	✓	✓	**56.04**	**60.64**

Table 2: Video emotion recognition accuracy (%) with different multimodal fusion modules

Fusion Strategy	VE-8	YF-6
Simple concatenation	50.27	57.09
Divide attention	52.20	58.56
Joint attention	54.12	60.27
Inner-modal only	54.40	58.44
Cross-modal only	46.15	48.29
Cross-temp multimodal	**56.04**	**60.64**

contribute to the performance of the cross-temporal multimodal fusion module.

To quantitatively verify the rationality of the emotional word predicting task, we compare the experimental results of different TSC masking strategies by: randomly masking 10%, 20%, or 30% of the words, and by masking all emotional words in each TSC. As shown in Table 3, as the number of masked words increases, the pre-trained representation performs more poorly on video emotion recognition tasks. That is because text semantics are reduced, so the pre-training task must emphasize text semantic understanding. The emotional word masking strategy yields the best performance on video emotion recognition, indicating that our word predicting task is emotion-oriented and effective in emotional representation learning.

As seen in Eq. 10, the trade-off of the two TSC-based pre-training tasks are weighted by two parameters λ_w and λ_t. We perform an ablation study on different ratios of λ_w and λ_t, experimental results are shown in Table 4. From this table, we observe that only using a single TSC-based pre-training task greatly reduces the performance of video emotion recognition. Since TSCs in different segments usually contain different emotional words, only predicting these emotional words based on the joint video feature will make it difficult for the video representation learning to converge. Without any emotional supervision, only aligning the video segment features with the TSCs will increase the variation of segment features. In addition, the method pretrained by two tasks obtains higher accuracy than the baseline. It shows that the combination of these two tasks can eliminate the convergence problem, and provide segment-level emotional supervision for video representation learning. Moreover, the performance of our method with λ_w:$\lambda_t = 1 : 1$ is higher than the settings of $1 : 2$ and $2 : 1$, which shows that the emotional word and appearing time predicting tasks contribute equally to affective video representation learning.

Table 3: Video emotion recognition accuracy (%) with different masking strategies

Masking Strategy	VE-8	YF-6
random 10%	53.85	53.55
random 20%	49.18	55.13
random 30%	45.33	50.61
emotional words	**57.53**	**61.00**

Table 4: Ablation study on emotional word predicting (EWP) and appearing time predicting (ATP) tasks, $\lambda_w : \lambda_t$ is the trade-off between these two tasks

Pre-training	VE-8	YF-6
No pre-training	56.04	60.64
EWP only	51.25	53.30
ATP only	42.36	50.73
$\lambda_w : \lambda_t = 1 : 2$	51.30	54.59
$\lambda_w : \lambda_t = 2 : 1$	52.29	59.50
$\lambda_w : \lambda_t = 1 : 1$	57.53	61.00

5.5 Comparisons to State-of-the-Art Methods

We compare our results to top methods on two user-generated video datasets, i.e., VE-8 and YF-6, as well as the movie dataset LIRIS-ACCEDE. The emotion recognition methods for user-generated videos include CFN [4], Kernelized [36], CSS [28], VAANet [38], Dual [20], ITE [27], KeyFrame [25], and FAEIL [37]. The affective analysis methods for movies include RBN [9], MMDRBN [23], MML [34], and AFRN [35]. These methods mainly perform affective analysis on one type of video. However, both user-generated videos and movies contain common affective patterns, so experiments are conducted on both kinds of video.

The top of Table 5 indicates that the proposed method without TSC-based pre-training (TAM w/o TSC) performs well on the VE-8 and YF-6 datasets. The proposed method achieves 56.04% and 60.64% accuracy respectively, which is 1.54% and 3.27% higher than the next best methods lacking auxiliary data. To analyze the performance gain of our method, we use the same features as the Dual [20] method, i.e. action, scene, and object features to perform emotion recognition on the VE-8 and YF-6 datasets. The result is shown in Table 5 TAM† w/o TSC. From this result, when using the same features as the Dual method, our method obtains 0.91% and 1.55% accuracy higher than the Dual method on VE-8 and YF-6 datasets. The Dual method uses the attention module and LSTM to learn the spatial and temporal dependencies, while our cross-temporal multimodal fusion module uses the self-attention module to simultaneously learn cross-temporal and cross-modal dependencies. The result shows that our fusion module has better capability than the Dual method. In addition, when using the motion, image, and audio features, our method gets 1.79% and 1.72% accuracy improvement than using Dual features on VE-8 and YF-6 datasets. Therefore, the motion, image, and audio features are more suitable for affective

Table 5: Video emotion recognition accuracy (%) comparison with other methods. *Auxiliary* indicates if other datasets are used for training. *TAM w/o TSC* indicates the proposed TAM method without TSC-based pre-training. *TAM† w/o TSC* represents our method using the same features like the Dual method.

Method	Auxiliary	VE-8	YF-6
CFN [4]		50.60	51.80
Kernelized [36]		52.50	54.40
CSS [28]		51.48	55.62
VAANet [38]		54.50	55.30
Dual [20]		53.34	57.37
TAM† w/o TSC		54.25	58.92
TAM w/o TSC		**56.04**	**60.64**
ITE [27]	✓	52.60	51.20
KeyFrame [25]	✓	52.85	59.51
FAEIL [37]	✓	**57.63**	60.44
TAM	✓	57.53	**61.00**

Table 6: Affective analysis results on the LIRIS-ACCEDE dataset compared to other methods. *TAM w/o TSC* stands for the proposed TAM method without TSC-based pre-training.

Methods	MediaEval2015		MediaEval2016			
	Valence	Arousal	Valence		Arousal	
	Acc	Acc	MSE	PCC	MSE	PCC
RBN [9]	44.26	64.30	0.332	0.387	0.766	0.416
MMDRBN [23]	46.73	65.10	0.303	0.450	0.713	0.470
MML [34]	46.22	57.40	0.198	0.399	1.173	0.446
AFRN [35]	48.61	58.22	0.193	0.468	**0.524**	0.522
TAM w/o TSC	**49.33**	**65.53**	**0.172**	**0.529**	1.115	**0.550**
TAM	**50.18**	**66.31**	**0.177**	**0.533**	0.754	**0.560**

video content analysis than the features used by the Dual method. In summary, the performance gain of our method is from both the feature choices and model designs.

Table 6 shows the effectiveness of our method on the affective movie analysis. Specifically, on the MediaEval2015 classification task, our method achieves an accuracy of 49.33% for valence and 65.53% for arousal, which is 0.72% and 0.43% higher than the next best methods. On the MediaEval2016 regression task, the MSE is 0.172 for valence and 1.115 for arousal, and the PCC is 0.529 for valence and 0.550 for arousal. With the exception of the MSE in arousal regression, all other results are significantly improved. One possible reason for the inferior performance of MSE is the impact of data distribution. Dellandréa et al. [6] show that MSE is not always sufficient to analyze models' efficiency when a large portion of the data is closed to a constant. For the arousal values of the MediaEval2016 test set, 45% are distributed in [3, 3.4] and 32% are distributed in [4.2, 4.6]. A model that always outputs 3.3 will result in low MSE performance, i.e. 0.6159. However, it does not mean this model predicts arousal values well. In this case, PCC is more suitable as the model evaluation metric than MSE. Thus, we use the PCC as the model evaluation metric to choose the best model during training. As a result, our model tends to learn better correlations between arousal values, resulting in better PCC at the expense of MSE on the MediaEval2016 test set. In summary, our TAM without TSC-based pre-training outperforms nearly all state-of-the-art methods on every task, including some methods with auxiliary data. Our method surpasses the simple fusion methods [28, 38] and cross-modal fusion methods [9, 20, 23, 35], showing the effectiveness of our cross-temporal multimodal fusion module on affective video content analysis.

As shown in Table 5, our temporal-aware multimodal method with TSC-based pre-training (TAM) achieves 57.53% and 61.00% accuracy on VE-8 and YF-6. These scores are 1.49% and 0.36% higher compared to the TAM without TSC-based pre-training. Compared to other methods using auxiliary emotion-labeled data [25, 27, 37], our pre-training framework performs well without any additional emotional labels. It also shows improvement on both the MediaEval2015 classification task and the MediaEval2016 regression task after TSC-based pre-training, as shown in Table 6. After TSC-based pre-training, the TAM improves valence and arousal accuracies by 0.85% and 0.78% on the MediaEval2015 task. On the MediaEval2016 task, the TAM after TSC-based pre-training improves valence PCC and arousal PCC by 0.004 and 0.01, and reduces arousal MSE by 0.361. These results demonstrate that TSC-based pre-training can effectively enhance representation learning for affective video content analysis.

6 CONCLUSION

In this work, we propose an effective temporal-aware multimodal method for affective video content analysis. This TAM method consists of cross-temporal multimodal fusion and TSC-based pre-training. The cross-temporal multimodal fusion module employs attention-based fusion to different modalities within each segment and across different segments. As a result, it fully captures the temporal dependencies between visual and audio signals. The experimental results show that the fusion module can mine more temporal patterns across all modalities for affective video content analysis. Moreover, we leverage TSCs to pre-train the fusion module for temporal-level representation learning, since TSCs are easily accessible and contain affective cues. Two self-supervised tasks are used to pre-train the fusion module. The emotional word predicting task predicts the emotional words in a TSC under the guidance of video representation and TSC semantics. The appearing time predicting task aims to predict when the TSC appears by calculating the similarities between video representation and TSC embedding. These pre-training tasks successfully mine the emotional cues and these cues as temporal-level supervision for representation learning. Comparison experiments verify that the cross-temporal multimodal fusion module can learn more discriminative representation after TSC-based pre-training.

ACKNOWLEDGMENTS

This work was supported by National Natural Science Foundation of China 92048203 and project from Anhui Science and Technology Agency 202104h04020011.

REFERENCES

[1] Esra Acar, Frank Hopfgartner, and Sahin Albayrak. 2014. Understanding affective content of music videos through learned representations. In *International conference on multimedia modeling*. Springer, 303–314.

[2] Yoann Baveye, Emmanuel Dellandrea, Christel Chamaret, and Liming Chen. 2015. LIRIS-ACCEDE: A video database for affective content analysis. *IEEE Transactions on Affective Computing* 6, 1 (2015), 43–55.

[3] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is space-time attention all you need for video understanding. *arXiv preprint arXiv:2102.05095* 2, 3 (2021), 4.

[4] Chen Chen, Zuxuan Wu, and Yu-Gang Jiang. 2016. Emotion in context: Deep semantic feature fusion for video emotion recognition. In *Proceedings of the 24th ACM international conference on Multimedia.* 127–131.

[5] Xu Chen, Yongfeng Zhang, Qingyao Ai, Hongteng Xu, Junchi Yan, and Zheng Qin. 2017. Personalized key frame recommendation. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval.* 315–324.

[6] Emmanuel Dellandréa, Liming Chen, Yoann Baveye, Mats Viktor Sjöberg, and Christel Chamaret. 2016. The mediaeval 2016 emotional impact of movies task. In *CEUR Workshop Proceedings*.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[9] Quan Gan, Shangfei Wang, Longfei Hao, and Qiang Ji. 2017. A multimodal deep regression bayesian network for affective video content analyses. In *Proceedings of the IEEE International Conference on Computer Vision.* 5113–5122.

[10] Xun Gao, Yin Zhao, Jie Zhang, and Longjun Cai. 2021. Pairwise Emotional Relationship Recognition in Drama Videos: Dataset and Benchmark. In *Proceedings of the 29th ACM International Conference on Multimedia.* 3380–3389.

[11] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. 2018. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition.* 6546–6555.

[12] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. 2017. CNN architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 131–135.

[13] Yu-Gang Jiang, Baohan Xu, and Xiangyang Xue. 2014. Predicting emotions in user-generated videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 28.

[14] Chenchen Li, Jialin Wang, Hongwei Wang, Miao Zhao, Wenjie Li, and Xiaotie Deng. 2019. Visual-texual emotion analysis with deep coupled video and danmu neural networks. *IEEE Transactions on Multimedia* 22, 6 (2019), 1634–1646.

[15] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557* (2019).

[16] Tao Liang, Guosheng Lin, Lei Feng, Yan Zhang, and Fengmao Lv. 2021. Attention Is Not Enough: Mitigating the Distribution Discrepancy in Asynchronous Multimodal Sequence Fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 8148–8156.

[17] Guangyi Lv, Tong Xu, Enhong Chen, Qi Liu, and Yi Zheng. 2016. Reading the videos: Temporal labeling for crowdsourced time-sync videos based on semantic embedding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 30.

[18] Trisha Mittal, Puneet Mathur, Aniket Bera, and Dinesh Manocha. 2021. Affect2mm: Affective analysis of multimedia content using emotion causality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 5661–5671.

[19] Fan Qi, Xiaoshan Yang, and Changsheng Xu. 2021. Zero-shot Video Emotion Recognition via Multimodal Protagonist-aware Transformer Network. In *Proceedings of the 29th ACM International Conference on Multimedia.* 1074–1083.

[20] Haonan Qiu, Liang He, and Feng Wang. 2020. Dual Focus Attention Network For Video Emotion Recognition. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1–6.

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, 8748–8763.

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[23] Shangfei Wang, Longfei Hao, and Qiang Ji. 2019. Knowledge-augmented multimodal deep regression bayesian networks for emotion video tagging. *IEEE Transactions on Multimedia* 22, 4 (2019), 1084–1097.

[24] Shangfei Wang and Qiang Ji. 2015. Video affective content analysis: a survey of state-of-the-art methods. *IEEE Transactions on Affective Computing* 6, 4 (2015), 410–430.

[25] Jie Wei, Xinyu Yang, and Yizhuo Dong. 2021. User-generated video emotion recognition based on key frames. *Multimedia Tools and Applications* 80, 9 (2021), 14343–14361.

[26] Bin Wu, Erheng Zhong, Ben Tan, Andrew Horner, and Qiang Yang. 2014. Crowdsourced time-sync video tagging using temporal and personalized topic modeling. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining.* 721–730.

[27] Baohan Xu, Yanwei Fu, Yu-Gang Jiang, Boyang Li, and Leonid Sigal. 2016. Heterogeneous knowledge transfer in video emotion recognition, attribution and summarization. *IEEE Transactions on Affective Computing* 9, 2 (2016), 255–270.

[28] Baohan Xu, Yingbin Zheng, Hao Ye, Caili Wu, Heng Wang, and Gufei Sun. 2019. Video emotion recognition with concept selection. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 406–411.

[29] Linhong Xu, Hongfei Lin, Yu Pan, Hui Ren, and Jianmei Chen. 2008. Constructing the affective lexicon ontology. *Journal of the China society for scientific and technical information* 27, 2 (2008), 180–185.

[30] Linli Xu and Chao Zhang. 2017. Bridging video content and comments: Synchronized video description with temporal summarization of crowdsourced time-sync comments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 31.

[31] Liang Yang and Hongfei Lin. 2012. Construction and application of Chinese emotional corpus. In *Workshop on Chinese Lexical Semantics*. Springer, 122–133.

[32] Wenmian Yang, Wenyuan Gao, Xiaojie Zhou, Weijia Jia, Shaohua Zhang, and Yutao Luo. 2019. Herding effect based attention for personalized time-sync video recommendation. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 454–459.

[33] Wenmain Yang, Kun Wang, Na Ruan, Wenyuan Gao, Weijia Jia, Wei Zhao, Nan Liu, and Yunyong Zhang. 2019. Time-sync Video Tag Extraction Using Semantic Association Graph. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 13, 4 (2019), 1–24.

[34] Yun Yi and Hanli Wang. 2019. Multi-modal learning for affective content analysis in movies. *Multimedia Tools and Applications* 78, 10 (2019), 13331–13350.

[35] Yun Yi, Hanli Wang, and Qinyu Li. 2019. Affective video content analysis with adaptive fusion recurrent network. *IEEE Transactions on Multimedia* 22, 9 (2019), 2454–2466.

[36] Haimin Zhang and Min Xu. 2018. Recognition of emotions in user-generated videos with kernelized features. *IEEE Transactions on Multimedia* 20, 10 (2018), 2824–2835.

[37] Haimin Zhang and Min Xu. 2021. Recognition of Emotions in User-generated Videos with Transferred Emotion Intensity Learning. *IEEE Transactions on Multimedia* (2021).

[38] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 303–311.

[39] Sicheng Zhao, Shangfei Wang, Mohammad Soleymani, Dhiraj Joshi, and Qiang Ji. 2019. Affective computing for large-scale heterogeneous multimedia data: A survey. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 15, 3s (2019), 1–32.