AVES: An Audio-Visual Emotion Stream Dataset for Temporal Emotion Detection

Yan Li, Wei Gan, Ke Lu, Dongmei Jiang*, and Ramesh Jain, *Fellow, IEEE*

Abstract—Human emotions vary over time, which can be vividly described as a stream of emotions. Observing the emotion stream in daily life provides valuable insights into an individual's mental state. However, existing research in emotion understanding has mainly focused on classification tasks, assigning an emotion category to a well-trimmed segment or each frame within a continuous signal. In contrast, the task of temporal emotion detection, which involves *locating* the boundaries of emotion segments and *recognizing* their categories in untrimmed signals, has not been fully explored. To advance research in this area, this paper introduces an in-the-wild Audio-Visual Emotion Stream (AVES) dataset, which is reliably annotated with the time boundaries and emotion category for each emotion segment in the videos. Thus, AVES can serve as a solid benchmark for temporal emotion detection tasks. Moreover, considering the flexible boundaries and varying durations of emotion segments, we propose a Boundary Combination Network (BoCoNet) for temporal emotion detection, which leverages short-term temporal context information to first predict the boundaries of emotion segments and then locate the entire emotion segments. Extensive experiments conducted on various representative unimodal and multimodal representations demonstrate that BoCoNet achieves state-of-the-art results. The AVES dataset will be released to the research community. We expect that this paper can advance the research on emotion stream and temporal emotion detection.

Index Terms—Temporal emotion detection, emotion stream, dataset, multimodal emotion recognition

1 Introduction

EMOTIONS play a crucial role in human daily life, shaping the processes of thinking, communicating, and decision-making [1]. Especially, human emotions are subject to change due to external stimuli and internal factors [2], [3]. Detecting changes in emotions from continuous signals, i.e. emotion stream, can enable human-computer interaction systems to provide more personalized responses. More importantly, long-term continuous observation of the emotion stream in a person's daily life not only aids in the prevention, early detection, precise evaluation, and prognostic management of mental health issues, but also assists the individual in increasing the awareness and proactive self-regulation of emotional well-being [4].

However, existing researches on emotion understanding primarily focus on classification tasks, assigning an emotion category to a well-trimmed segment [5], [6], [7] or each frame of a continuous signal [8], [9], [10]. In our previous work [11], we first introduced the task of temporal emotion detection, which aims to detect all emotion segments within an untrimmed, continuous signal. Unlike common frame-level, image-level, and utterance-level emotion classification tasks, temporal emotion detection requires not only localizing the temporal boundaries, but also predicting the corresponding emotion category, of each emotion segment. This task takes into consideration the continuity and completeness of emotion segments, and is more aligned with the real-world scenarios of people experiencing emotions.

To facilitate the research on temporal emotion detection, an emotion stream dataset annotated with the temporal boundaries and emotion category of each emotion segment along untrimmed signals is crucial. However, most existing emotion datasets are not suitable for this challenging task. On one hand, some datasets have been collected for utterance-level emotion recognition tasks, where each sample was well-trimmed and annotated with an emotion category. These datasets include RML [12], eNTERFACE [13], and IEMOCAP [14] which were recorded in laboratory environments, as well as MELD [15], DFEW [16], and FERV39k [17] which were collected from online videos or TV series. On the other hand, some datasets have been released for frame-level or image-level emotion recognition tasks. For example, the Aff-wild2 dataset [18] annotates each video frame with specific facial expression categories. Similarly, the EmoSet dataset [19] labels images, ranging from social to artistic contexts, with various emotion categories. These datasets ignore the continuity and completeness of emotion segments in real-life scenarios and therefore can not be used for the temporal emotion detection tasks.

Several datasets are closely related to the task of temporal emotion detection. For example, CAS(ME)^2 [20] and

* *Corresponding author*
- *Y. Li is with the Shaanxi Key Laboratory on Speech and Image Information Processing, National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, School of Computer Science, Northwestern Polytechnical University (NPU), Youyi Xilu 127, Xi'an 710072, China. He is also an intern student in the Peng Cheng Laboratory, No.2 Xingke 1st Street, Nanshan District, Shenzhen 518055, China. E-mail: liyan4ai@gmail.com*
- *D. Jiang is with the Peng Cheng Laboratory, No.2 Xingke 1st Street, Nanshan District, Shenzhen 518055, China. She is also with the School of Computer Science, Northwestern Polytechnical University (NPU), Youyi Xilu 127, Xi'an 710072, China. E-mail: jiangdm@pcl.ac.cn*
- *W. Gan and K. Lu are with the School of Engineering Science, University of Chinese Academy of Sciences, No.19A Yuquan Road, Beijing 100049, China. They are also with the Peng Cheng Laboratory, No.2 Xingke 1st Street, Nanshan District, Shenzhen 518055, China. E-mail: ganwei19@mails.ucas.ac.cn, luk@ucas.ac.cn*
- *R. Jain is with the Department of Computer Science, University of California, Irvine. E-mail: jain@ics.uci.edu*

Manuscript received December 2023.

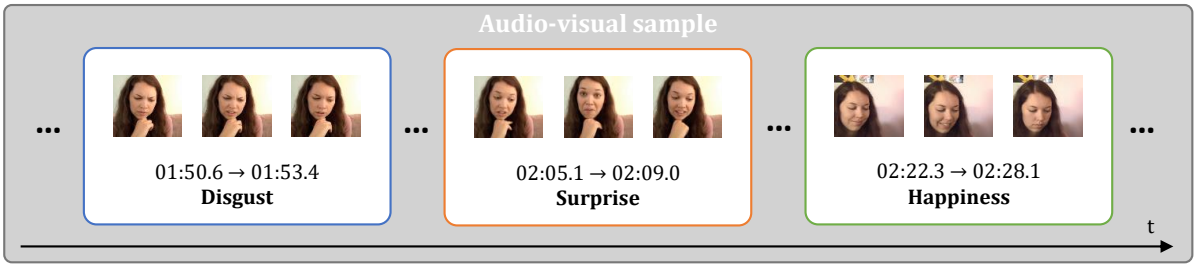


Fig. 1. An audio-visual emotion stream sample with three emotion segments in the AVES dataset. In contrast to traditional research that aims at classifying emotions from pre-trimmed segments, our study focuses on detecting emotion segments from continuous signals.

CAS(ME)^3 datasets [21] were recorded in the laboratory environment for the micro-expression and macro-expression spotting tasks. Both datasets were annotated with micro-expression segments lasting less than 0.5 seconds and macro-expression segments lasting between 0.5 and 4 seconds within facial videos. Unfortunately, these datasets do not account for the in-the-wild environment and the audio modality. The authors of [22] developed the TSL-300 sentiment dataset by combining and re-annotating the raw videos from multiple existing datasets, including CMU-MOSEI [23], CMU-MOSI [24], Ekman6 [25], and VideoEmotion8 [26]. Four annotators were recruited to label the emotion segments in these videos with two sentiments: positive and negative. The SDFE-LV dataset [27] was collected for spotting dynamic facial expressions in long videos, where five annotators marked the time boundaries and emotion categories of expression segments within each long video. However, the SDFE-LV dataset suffers from severe data imbalance, with only two samples of the fear category, and does not take into account the audio modality. In our previous research [11], the MEMOS dataset was recorded in the laboratory environment, where multiple participants played the Monopoly game together to induce changes of emotion. Although each multimodal recording was self-annotated with the emotion segments by the participants, the reliability of these annotations does not sufficiently make MEMOS a solid testbed for temporal emotion detection. Overall, it is imperative to collect and annotate a reliable multimodal emotion stream dataset in the wild to boost the new research area on temporal emotion detection.

In this paper, we present the Audio-Visual Emotion Stream (AVES) dataset for the task of temporal emotion detection, which was collected from online websites and contains rich emotional changes in wild environments. Each audio-visual sample was annotated in the form of an emotion stream, including the starting time, ending time, and emotion category for each emotion segment, as shown in Fig. 1. To ensure the reliability of the annotations, each sample was labeled by seven expert annotators, and samples with low annotation consistency were discarded. Finally, the AVES dataset comprises a total of 167 audio-visual samples lasting approximately 13 hours, in which 1,747 emotion segments are labeled covering six basic emotions. Compared to the existing emotion datasets, AVES offers advantages in terms of novelty of the task, challenges of the samples, and reliability of the annotations.

Apart from the emotion stream datasets, current research on temporal emotion detection methods is quite limited. A straightforward approach is to predict the emotion category at each moment and then merge adjacent frames with the same category to generate emotion segments. Our previous research [11] introduced a method that utilized sliding windows and a binary classifier to detect emotion segments. However, the predefined sliding window constrains the position and duration of detected emotion segments, making the detected temporal boundaries not precise. For the task of dynamic facial expression spotting, the authors of [27] utilized the Boundary-Matching Network to predict the confidence of facial expression segments, which fails to take into account the completeness of emotion segments. Moreover, for micro-expression and macro-expression spotting tasks, beyond methods based on binary classifiers [28], some researchers have adopted optical flow features to calculate the differences within windows and locate micro- and macro-expression events [29], [30]. The corresponding datasets, collected in laboratory environments, allow optical flow feature variations to effectively reflect subtle facial movements. However, such methods lack robustness in wild environments with complex lighting and head movements.

In this paper, we propose a Boundary Combination Network (BoCoNet) for temporal emotion detection. Specifically, the proposal generation module utilizes short-term temporal context information to predict the boundaries of emotion segments and combines them to generate emotion proposals. Compared to the pre-defined window methods, our method can cover various boundaries and durations of emotion segments. Subsequently, the completeness discrimination module employs high-level temporal context information to assess the completeness of emotion segments and remove those that are incomplete, further improving the reliability of detection results.

In summary, the contributions of this paper are in three folds:

- We collect an Audio-Visual Emotion Stream (AVES) dataset, which provides a reliable yet challenging benchmark for temporal emotion detection tasks.
- We propose a Boundary Combination Network (BoCoNet) for temporal emotion detection, which can detect emotion segments with flexible durations and precise boundaries in a continuous signal.
- Extensive experiments and analyses are conducted on the AVES dataset. Results demonstrate that the proposed BoCoNet achieves state-of-the-art performance with various representation features.

TABLE 1
Comparison of AVES with related sentiment, emotion, and expression datasets.

Task	Dataset	Wild	Annotation	Audio	Annotator	Category†	Sample‡	Limitation	Year
cls.	RML [12]	✗	utterance	✓	-	6	500	Datasets for traditional emotion or expression classification tasks.	2008
	IEMOCAP [14]	✗	utterance	✓	3	10	7,433		2008
	eNTERFACE [13]	✗	utterance	✓	2	6	1,166		2006
	AFEW [31]	✓	utterance	✓	2	7	1,747		2012
	DFEW [16]	✓	utterance	✓	10	7	16,372		2020
	MELD [15]	✓	utterance	✓	3	7	13,708		2018
	FERV39k [17]	✓	utterance	✓	30	7	39,546		2022
	CMU-MOSEI [23]	✓	utterance	✓	3	6	23,453		2018
	Aff-wild2 [18]	✓	frame	✓	3	7	403,758		2018
	EmoSet [19]	✓	frame	✗	10	8	118,102		2023
det.	SMIC-E-Long [32]	✗	stream	✗	2	3+1	167	ME spotting task.	2021
	CAS(ME)^2 [20]	✗	stream	✗	2	4+1	247	ME spotting task.	2017
	CAS(ME)^3 [21]	✗	stream	✗	2	7+1	4599	ME and MaE spotting tasks.	2022
	MEMOS [11]	✗	stream	✓	1	5+1	6,317	Limited annotation reliability.	2020
	SDFE-LV [27]	✓	stream	✗	5	6+1	2420	Unimodal dataset.	2022
	TSL-300 [22]	✓	stream	✓	4	2+1	1,642	Limited sentiment categories.	2022
	AVES	✓	**stream**	✓	**7**	**7+1**	**1747**		**2023**

† To align with classification datasets, we include the background category (i.e., neutral) of detection datasets.
‡ It is unfair to compare the sample size between classification and detection datasets.

The remainder of this paper is organized as follows. We begin by reviewing the related work in Section 2. In Section 3, we provide a detailed description of the AVES dataset. Section 4 introduces the proposed Boundary Combination Network (BoCoNet). Experimental results and analysis are presented in Section 5. Ethical considerations and conclusions are discussed in Section 6.

2 Related Work

In this section, we review the related work in terms of both datasets and methods.

2.1 Emotion and Expression Datasets

Existing emotion recognition datasets can be divided into two categories according to their annotations, as shown in Table 1. One category is the *utterance-level* datasets. For example, in the RML [12] and eNTERFACE [13] datasets, each participant is induced or instructed to express a specific emotion. IEMOCAP [14] is also an utterance-level dataset, where ten actors in dyadic sessions perform selected emotional scripts and improvised hypothetical scenarios designed to elicit specific emotions. RML, eNTERFACE, and IEMOCAP are recorded in a laboratory environment, while AFEW [31], DFEW [16], MELD [15], FERV39k [17], and CMU-MOSEI [23] source their data from movies, TV shows, or online video websites. These utterance-level datasets have been manually trimmed to ensure that each sample contains only one emotion, which fails to capture the emotional changes that occur in real-life scenarios. The other category is *image-level* or *frame-level* datasets. For example, the EmoSet dataset [19] annotates each social or artistic image with eight emotion categories and six emotion attributes. Similarly, the Aff-wild2 dataset [18] is collected from online video websites and annotated with frame-level facial expression categories. However, it disregards the continuity and completeness of emotional expression and therefore is not suitable for temporal emotion detection tasks.

In addition to emotion recognition datasets, there are also some micro-expression and macro-expression spotting datasets. Common in-the-lab expression spotting datasets include SMIC-E-Long [32], CAS(ME)^2 [20], and CAS(ME)^3 [21]. In these datasets, micro-expressions lasting less than 0.5 seconds and macro-expressions ranging from 0.5 to 4 seconds are annotated. SDFE-LV [27], an in-the-wild dynamic facial expression spotting dataset, accounts for macro-expression segments longer than 4 seconds. It is important to note that the emotion categories in SDFE-LV are severely imbalanced, containing only two Fear samples. More importantly, these datasets focus on locating expressions through visual modalities and do not consider the auditory modality, which is indispensable when human expresses emotions.

MEMOS [11] is the first multimodal dataset for the task of temporal emotion detection. It records the time boundaries of emotion segments and their corresponding emotion categories for each participant while playing the Monopoly game. Unfortunately, the self-annotation is not sufficiently reliable, and MEMOS does not contain samples of fear and disgust. These two limitations hinder MEMOS as a solid benchmark for temporal emotion detection tasks. Similarly, the recent TSL-300 dataset [22] is labeled by four annotators and only includes two sentiments: positive and negative. In contrast, the proposed AVES dataset is collected from online video websites and contains rich emotional changes covering six basic emotions. Moreover, each video in AVES is labeled by seven expert annotators to ensure the high reliability of the annotations.

2.2 Emotion Detection and Expression Spotting Methods

Currently, research on methods for temporal emotion detection is relatively limited. In our earlier study [11], a binary classifier is utilized on sliding windows to predict potential emotion segments. Zhang et al. [22] first obtain frame-level sentiment scores, then use predefined thresholds and post-

processing to detect segments for each sentiment. Since both expression spotting tasks and temporal emotion detection tasks require locating temporal segments in continuous signals, here we introduce relevant expression spotting methods for a more comprehensive comparison.

In recent years, there has been some progress in research on expression spotting. Tran et al. [28] tackled micro-expression spotting as a binary classification problem based on a window sliding across positions and scales in an image sequence. Some hand-crafted features and a linear Support Vector Machine (SVM) were used to classify whether the current window is a micro-expression event or not. In [33], a Long Short-Term Memory (LSTM) network was utilized to predict the micro-expression apex frame score at each moment, followed by several heuristics rules to determine the positions of the micro-expression apex and spot micro-expression events. Similarly, the authors of [34] first predicted the confidence scores of micro-expressions at each moment. After smoothing the confidence, a threshold and peak detection technique was used to localize the micro-expressions. Furthermore, some methods locate micro-expression events by analyzing feature differences. Li et al. [35] adopted the Chi-Square distance to measure the differences between the Local Binary Pattern (LBP) features of sequential video frames within a specified interval, and applied a threshold and peak detection method to locate the peaks indicating the highest intensity frames of fast facial movements. The top two of the Micro-Expression Grand Challenge (MEGC) 2022 [29], [30], [36] adopted a similar method of feature difference comparison, where the micro-expressions were spotted by calculating the difference of optical flow features in a sliding window. For dynamic facial expression spotting tasks, Xu et al. [27] evaluated the effectiveness of the Boundary-Matching Network (BMN), which leverages a complex boundary-matching mechanism to predict the confidence of facial expression segments. However, the completeness of expression segments is not considered, which may lead to incomplete detection results.

It is important to note that there are significant differences between micro-expression spotting tasks and temporal emotion detection tasks. On one hand, micro-expression spotting focuses on capturing subtle and fast facial movements within a sequence of face images, while the goal of temporal emotion detection is to locate macroscopic emotional segments within continuous multimodal signals. On the other hand, micro-expression spotting datasets are typically collected in laboratory environments, with sample durations less than a minute [20], [37], [38]. In contrast, temporal emotion detection datasets are closer to in-the-wild environments, with the duration of samples that can last several minutes or even tens of minutes, which is more challenging. Therefore, it is necessary to design specialized networks for temporal emotion detection tasks.

3 AVES Dataset

In this section, we first introduce the data collection and preprocessing strategy, then describe the data annotation protocol in detail. Statistics on the number and duration of the emotion segments will be given. Finally, we summarize the highlights of the AVES dataset.

3.1 Data Collection and Preprocessing

Online video sites provide an opportunity for acquiring large amounts of data from various people and scenarios. The proposed AVES dataset is constructed by videos from two online video sites, namely YouTube and Bilibili. Scenarios that contain rich changes in emotion are considered, for instance, reactions to movie trailers, comments on current social events, and sharing of personal lives. The keywords used to retrieve the videos are "reaction" and other emotion-related words from Plutchik's Wheel of Emotions [39]. The videos selected are limited to setups where the subject's attention is exclusively towards the camera. Videos with moving cameras, such as cameras on bikes or selfie recordings while walking, are discarded. Fig. 2 shows a typical and recommended scenario, the reaction to a movie trailer. Movie trailers can induce changes in the subject's emotions, and different types of trailers can stimulate different emotions.

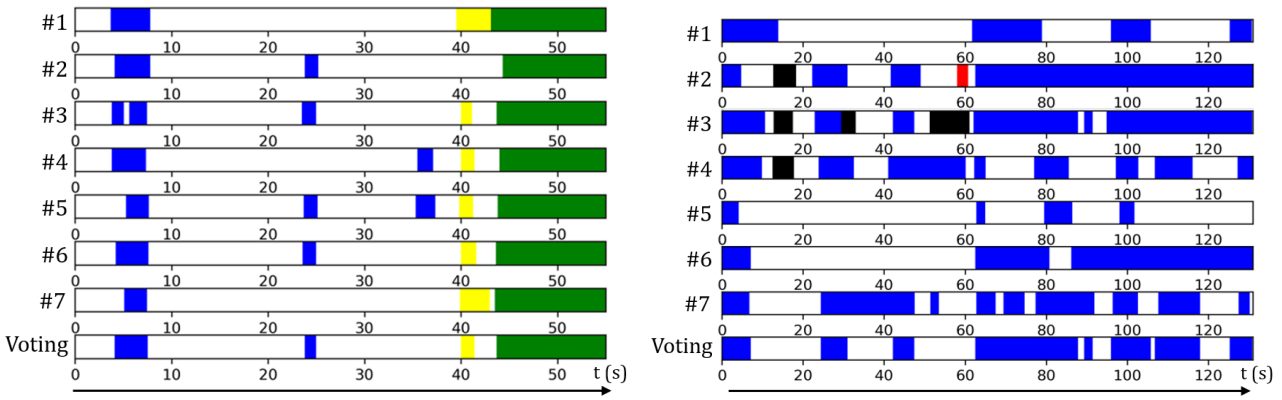
Fig. 2. The movie trailer reaction scenario.

The videos are downloaded at the highest available resolution. In cases where a video contains more than one individual, we split it into multiple videos, each containing only one individual. Therefore, the resolutions of the videos are not uniform. We manually select the long videos with rich emotional changes. As a result, 174 video samples are obtained. These videos are further transcoded into MPEG4 format and re-sampled to 30fps. The corresponding audio is standardized to a sampling rate of 16kHz. It should be noted that the audio may contain some background noise.

3.2 Dataset Annotation

An emotion stream includes the starting time, ending time, and emotion category of each emotional segment within a continuous signal. Each annotator first looks through the entire video to understand its content and then repeatedly watches the video to precisely locate the boundaries of each emotional segment. For each segment, annotators are asked to infer the emotion category from individual vocal, facial, and linguistic cues, with video stimuli regarded as contextual information to aid in the inference. Besides Ekman's six basic emotions, the "other" category is introduced to represent the complex emotions that do not fall into the six emotion categories. Once all the emotional segments have been labeled, the remaining parts of the video are considered as neutral.

Considering the subjective characteristic of emotions, seven graduate students with at least two years of research experience in affective computing are recruited and trained to annotate the videos. To ensure the quality of annotations, a senior annotator is responsible for reviewing their labels. Fig. 3 illustrates the emotion stream annotations of two videos, with different colors representing different emotions and white representing the neutral emotion.



(a) The annotation of a retained video. (b) The annotation of a discarded video.

Fig. 3. The annotation visualization of a retained video and a discarded video. The different colors represent different emotion categories, with white representing "neutral."

TABLE 2
The interpretation of Fleiss' kappa coefficient.

Fleiss' kappa coefficient	Interpretation
< 0	Poor agreement
0.01 - 0.20	Slight agreement
0.21 - 0.40	Fair agreement
0.41 - 0.60	Moderate agreement
0.61 - 0.80	Substantial agreement
0.81 - 1.00	Almost perfect agreement

We further discard the videos with low annotation consistency. Specifically, the Fleiss' kappa coefficient is adopted to evaluate the consistency of multiple annotations. Based on the explanation in Table 2, videos with a Fleiss' kappa coefficient below 0.2 are discarded to ensure the quality of the annotations. Fig. 3 (a) depicts a retained video with a Fleiss' kappa coefficient of 0.825. In contrast, Fig. 3 (b) shows a discarded video with a Fleiss' kappa coefficient of 0.165. In the end, 167 videos are retained in AVES, with a global Fleiss' kappa coefficient of 0.449, which is higher than the IEMOCAP dataset (0.4) [14], and the MELD dataset (0.43) [15]. Fig. 4 shows the distribution of Fleiss' kappa coefficients in AVES.

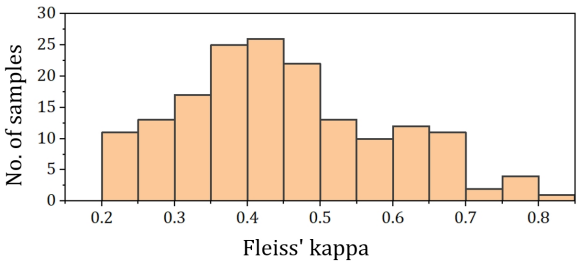


Fig. 4. Distribution of Fleiss' kappa coefficients in AVES.

The final emotion stream annotation is generated through majority voting on each frame. For segments with conflicting emotion categories, all annotators discuss together to reach an agreement on the emotion category based on contextual information. As a result, there are no compound emotions in AVES.

In general, we implement three measures to address potential biases during the annotation process: (1) We recruit seven annotators with at least two years of experience in affective computing research. (2) Each annotator undergoes training and trial annotation before the formal annotation begins. (3) Samples with low annotation consistency are discarded.

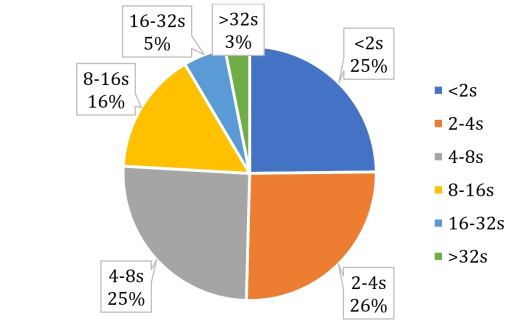
3.3 Dataset Statistics

As mentioned above, the AVES dataset consists of 167 videos with 1,747 audio-visual multimodal emotion segments. The duration of each video ranges from 40 seconds to 15 minutes, totaling approximately 13 hours. Some statistical information regarding the audio-visual multimodal emotion segments is illustrated in Fig. 5. Subfigure (a) displays the duration distribution of all emotion segments. It can be observed that the numbers of the segments shorter than 2 seconds, between 2 and 4 seconds, between 4 and 8 seconds, and longer than 8 seconds are quite close. A few segments are longer than 32 seconds. Subfigure (b) presents the number distribution of segments for each emotion. It is evident that "happiness" is the dominant emotion, followed by "surprise" and "sadness". "Fear" and "disgust" have a smaller number of samples, while "anger" and "other" emotions are rarely present in the dataset. Subfigure (c) illustrates the duration distribution of segments for each emotion. Compared to the other emotions, "sadness" and "anger" have longer segments. The length of each emotion segment varies greatly, which poses a significant challenge for temporal emotion detection tasks.

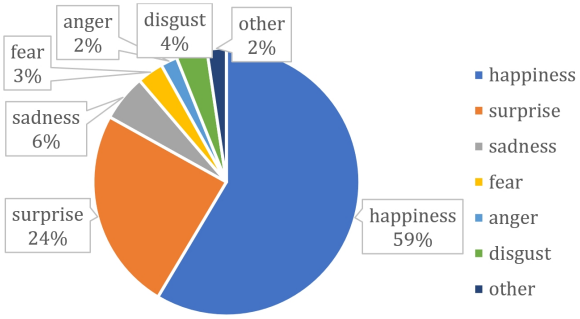
Following previous research [2], [24], [40], we split the AVES dataset by randomly selecting around 50% of the videos as the training set, 20% as the validation set, and the remaining 30% as the test set, while trying to keep the ratio of different emotion categories close in these subsets. The number of videos and the number of segments in the subsets are shown in Table 3.

3.4 Dataset Highlights

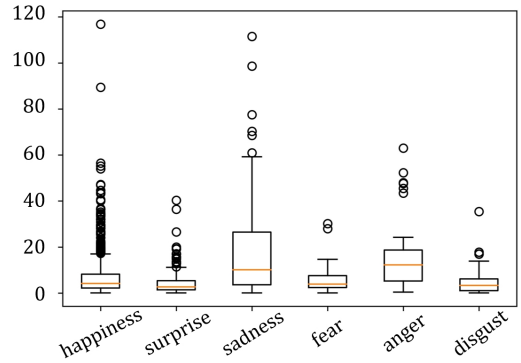
Compared with existing emotion datasets, the highlights of AVES can be summarized as follows:



(a) Duration of emotion segments.



(b) Number of segments for each emotion.



(c) Duration of segments for each emotion (in second).

Fig. 5. Statistics on emotion segments in AVES.

TABLE 3
Number of videos and emotion segments in each subset.

Subset	Duration (hh:mm)	Video	Segment
Train	05:36	73	767
Validation	02:35	35	337
Test	04:34	59	643
Total	12:45	167	1,747

- **Novelty.** Most existing emotion datasets are designed for utterance-level emotion recognition or frame-level facial expression recognition. In contrast, the AVES dataset annotates the untrimmed continuous multimodal signals with the emotion segments, which provides a benchmark for temporal emotion detection research.
- **Reliability.** Each video in AVES is labeled by seven annotators with research backgrounds in affective computing, and the videos with low annotation consistency are removed from AVES. This ensures that AVES can serve as a solid testbed for temporal emotion detection tasks.
- **Challenge.** Temporal emotion detection tasks are inherently challenging as they require not only recognizing the emotion categories but also localizing the boundaries of emotion segments. Moreover, the videos in the AVES dataset are in the wild, featuring complex environments and scenes, making the temporal task even more challenging.

4 BoCoNet for Temporal Emotion Detection

4.1 Problem Definition

An untrimmed signal is denoted as $X = \{x_t\}_{t=1}^{T}$, where T is the duration of X, and x_t is the multimodal representation at moment t. In the training set, X is annotated with a set of temporal emotion segments $\Phi = \{\phi_n = (\varphi_n, \varphi'_n, k_n)\}_{n=1}^{N}$, where N is the number of temporal emotion segments in X. φ_n, φ'_n, and k_n are the starting time, ending time, and category of the emotion segment ϕ_n, respectively. $k_n \in \{1, 2, \ldots, K\}$, where K is the number of emotion categories. The objective of the temporal emotion detection task is to determine Φ for a given untrimmed signal X.

4.2 Boundary Combination Network

4.2.1 Overview

The proposed Boundary Combination Network (BoCoNet) for temporal emotion detection is shown in Fig. 6. Firstly, an untrimmed video is transformed into a multimodal embedding sequence using a generalized multimodal encoder. Noting that our focus here is on the method for temporal emotion detection rather than multimodal representation learning, we adopt the pre-trained models such as the Two-Stream Aural-Visual (TSAV) network [41] or the Facepp toolkit [11] as the encoder. Based on the frame-level embedding sequence, the Proposal Generation Module utilizes temporal contextual information to construct the emotion segment proposals. Specifically, the short-term context is first used to predict potential boundaries of emotions, then the starting and ending boundaries are combined to generate the emotion segment proposals. The Emotion Recognition module is used to remove the neutral proposals. The retained non-neutral proposals are sent to the Completeness Discrimination module, where the emotional completeness of each proposal is further examined, and the incomplete proposals are discarded. Finally, the temporal emotion detection results are obtained by applying Soft Non-Maximum Suppression (Soft-NMS) [42] to remove the redundant and false-positive proposals.

4.2.2 Proposal Generation Module

The first and most important module of BoCoNet is the Proposal Generation Module consisting of two 1D convolutional transformer networks (ConvTrans1 and ConvTrans2), which are designed to predict whether the current moment is the starting or ending of any emotion segment, respectively. The convolutional transformer networks comprise two convolutional layers for short-term temporal context modeling, two transformer layers [43] for adaptive attention

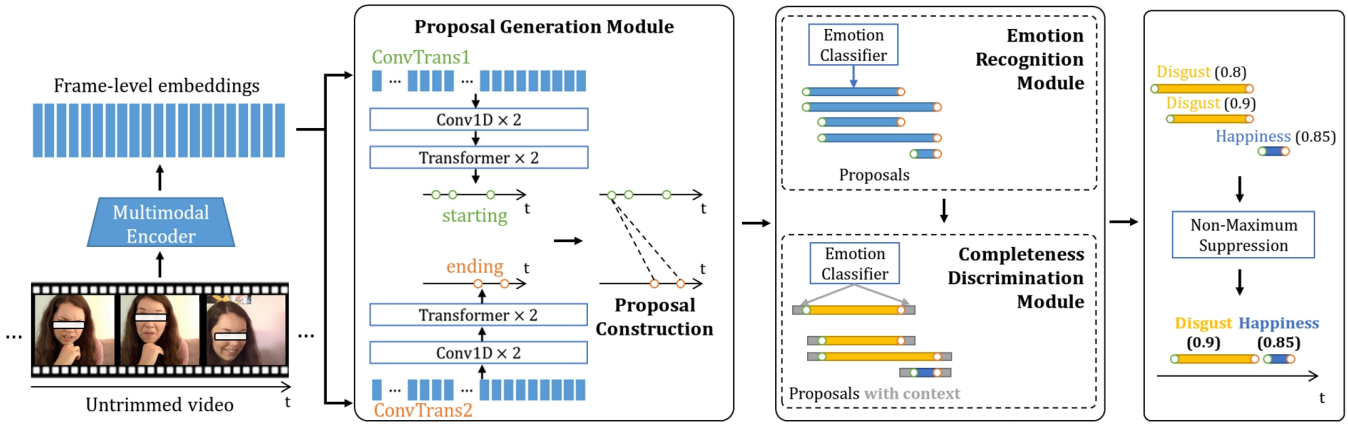


Fig. 6. The proposed boundary combination network (BoCoNet) for temporal emotion detection.

to emotionally salient moments and learning high-level representations, and a fully connected (FC) layer for prediction. The detailed architecture is illustrated in Fig. 7.

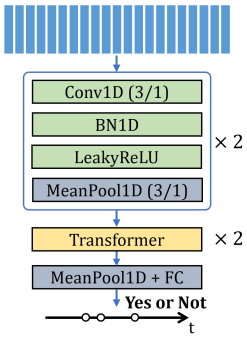


Fig. 7. The architecture of the convolutional transformer network, where the *kernel size* and *stride* of each layer are shown in the bracket.

Psychological studies [38], [44], [45] have indicated that facial movements lasting less than 500ms are considered as micro-expressions, while facial movements lasting longer than 500ms are considered as macro-expressions. Therefore, for each moment, ConvTrans utilizes surrounding one-second contextual information to predict whether the current moment is a boundary or not. It should be noted that a moment can simultaneously be both a starting boundary and an ending boundary. Once the potential starting and ending boundaries are determined, emotional segment proposals are constructed by combining all the starting and ending boundaries. Following the observation in the AVES dataset that seldom emotion segments last longer than 60 seconds, the length of proposals is limited to 60 seconds.

4.2.3 Emotion Recognition Module

The generated proposals are fed into the Emotion Recognition Module. For each proposal, average pooling is employed to integrate the frame-level embeddings to obtain a proposal-level representation, which is then fed into a pre-trained emotion classification model, i.e., a multi-layer perceptron trained on the emotion segments of the AVES training set, to recognize the emotion category and predict the confidence score. All the proposals classified as neutral are discarded directly, and only the non-neutral proposals will be retained.

4.2.4 Completeness Discrimination Module

Although the emotion recognition module can predict emotion categories, it cannot perceive the completeness of emotion segments. The incomplete emotion segment, as the false-negative sample, would reduce the detection precision. Therefore, we formulate a Completeness Discrimination Module. For each non-neutral proposal, two one-second context snippets (shown in grey at both ends in Fig. 6) are used to determine whether the proposal is complete or not. It is considered incomplete if the category of either of the two context snippets matches the category of the proposal. All incomplete proposals are discarded.

In short, a proposal is retained only if it meets the following two conditions: 1) The emotion category of the proposal is non-neutral; 2) The proposal is complete to express an emotion.

4.2.5 Soft-NMS

Considering that the detected emotional segments often overlap with each other, Non-Maximum Suppression (NMS) is employed to remove redundant and false-positive proposals. Conventional greedy NMS algorithms simply discard the segments with lower detection scores among the overlapping ones. In contrast, Soft-NMS [42] decreases the detection scores according to the degree of overlap (temporal Intersection over Union, tIoU). A high penalty is introduced when the overlap is high and a low penalty is introduced when the overlap is low, thus improving the detection precision. The detailed algorithm for Soft-NMS is presented in Algorithm 1.

5 Experiments

5.1 Multimodal Encoder

The following three representative encoders are employed in the experiment:

TSAV [41]: The Two-Stream Aural-Visual (TSAV) network is a multimodal emotion recognition model. For each moment, eight surrounding face images and the corresponding facial landmark masks are input to a 3D-ResNet

Input:
$\mathcal{S} = \{s_1, s_2, \ldots, s_N\}$, $\mathcal{C} = \{c_1, c_2, \ldots, c_N\}$, N_t.
$\mathcal{S}$ is the list of initial detection segments;
$\mathcal{C}$ contains corresponding detection confidences;
N_t is the soft-NMS threshold.
Output:
$\mathcal{D}$ is the list of final detection segments;
$\mathcal{C}$ contains updated detection confidences;
begin
  $\mathcal{D} \leftarrow \{\}$
  while $\mathcal{S} \neq$ *empty* **do**
    $m \leftarrow \text{argmax}\, \mathcal{C}$
    $\mathcal{M} \leftarrow s_m$;
    $\mathcal{D} \leftarrow \mathcal{D} \cup \mathcal{M}$
    $\mathcal{S} \leftarrow \mathcal{S} - \mathcal{M}$
    for $s_1 \in \mathcal{S}$ **do**
      /* Decrease confidence according to tIoU */
      $c_i \leftarrow c_i \times f(\text{tIoU}(\mathcal{M}, s_1))$
    endFor
  end
  return $\mathcal{D}, \mathcal{C}$
end

Algorithm 1: Soft Non-Maximum Suppression.

to learn a 512-dimensional visual representation, and a 10s speech spectrogram is fed to a 2D-ResNet to learn a 512-dimensional aural representation, resulting in a 1024-dimensional multimodal representation for each moment. Therefore, the TSAV feature contains not only information about the current moment but also temporal context information. Based on the TSAV model pre-trained on the Aff-wild2 dataset, we further fine-tune the model on the AVES training subset.

Data2vec [46]: Data2vec is a self-supervised representation learning framework. The motivation is to predict latent representations of the full input data based on a masked view of the input in a self-distillation setup using a standard Transformer architecture. We utilize a Data2vec-vision network pre-trained on ImageNet-1K [47] to extract a 768-dimensional frame-level visual feature, and a Data2vec-audio network pre-trained on Librispeech [48] to extract a 768-dimensional frame-level aural feature, and finally concatenate them as a 1,536-dimensional frame-level multimodal feature.

Facepp [11]: The Facepp expression recognition toolkit[1] is utilized to predict the confidence score for seven emotions (six basic emotions and neutral) of each facial image directly, which is used as the unimodal feature to generate the emotion segment proposals.

5.2 Experimental Setting

5.2.1 Implementation Details

For the backbone based on TASV, the pre-trained parameters are fine-tuned with a learning rate of 0.0001. The convolutional transformer network is trained for up to 10 epochs with a learning rate of 0.001. LeakyReLU with a negative slope of 0.2 is adopted as the non-linear activation function. Kaiming Initialization [49] is used to initialize network parameters. Adam [50] with a weight decay of 0.001 is utilized for network optimization. All the experiments are conducted on the PyTorch platform [51] using the TITAN RTX.

Considering the boundary and non-boundary classes are severely imbalanced, a soft-label strategy and the Focal loss [52] are adopted for training the convolutional transformer networks. Specifically, for ConvTrans1 and ConvTrans2, the moments within 0.5 seconds of the boundary of each emotion segment in the training set are regarded as positive samples, and the other moments are regarded as negative samples.

5.2.2 Evaluation Metrics

Inspired by other detection tasks, **Average-mAP** is adopted to evaluate the final detection results. It reflects the comprehensive detection precision under different recalls, different emotion categories, and different tIoU thresholds. Moreover, we further propose **Average-wAP** to consider the imbalance in emotion categories, and **Average-AP** to reflect the detection results of different emotions. In the experiments, we evaluate detection results on six basic emotions, and the tIoU thresholds of 0.1, 0.2, 0.3, 0.4, and 0.5 are adopted.

Average-mAP. For an emotion category $i \in \{1, 2, \ldots, I\}$ and a tIoU threshold $j \in \{1, 2, \ldots, J\}$, AP_{ij} is calculated to show the average detection precision under different recalls:

$$AP_{ij} = \int_{r=0}^{1} p_{ij}(r)dr, \tag{1}$$

where $p_{ij}(r)$ represents the recall-precision curve. The AP incorporates the trade-off between precision and recall, and considers both false positives (FP) and false negatives (FN). This property makes AP a suitable metric for most detection applications.

Then, mAP_j is computed to show the **unweighted** mean AP_{ij} across all emotion categories under the tIoU threshold j:

$$mAP_j = \frac{1}{I}\sum_{i=1}^{I} AP_{ij}. \tag{2}$$

Finally, the Average-mAP is calculated to show the average detection results across multiple thresholds:

$$\text{Average-mAP} = \frac{1}{J}\sum_{j=1}^{J} AP_j. \tag{3}$$

Average-wAP. Assuming that the number of samples of emotion i is n_j, the weight of emotion i is calculated as:

$$w_i = \frac{n_i}{\sum_{i=1}^{I} n_i}. \tag{4}$$

Then, wAP_j is computed to show the **weighted** mean AP_{ij} across all emotion categories under the tIoU threshold j:

$$wAP_j = \frac{1}{I}\sum_{i=1}^{I} w_i AP_{ij}. \tag{5}$$

Finally, the Average-mAP is calculated to show the average detection results across multiple thresholds:

$$\text{Average-wAP} = \frac{1}{J}\sum_{j=1}^{J} wAP_j. \tag{6}$$

1. https://console.faceplusplus.com/documents/6329465

TABLE 4
Temporal emotion detection results (%) in AVES.

Methods	TSAV		Data2vec		Facepp	
	Average-mAP	Average-wAP	Average-mAP	Average-wAP	Average-mAP	Average-wAP
Feature Difference Analysis [35]	< 1	< 1	< 1	< 1	< 1	< 1
Optical Flow Magnitude [29]	< 1	< 1	< 1	< 1	< 1	< 1
Pyramid Sliding Window [28]	9.22	15.76	3.98	6.33	6.47	14.20
VideoMAE-OF [53]	11.54	23.13	1.52	3.45	2.96	5.61
Boundary Matching Network [27]	12.79	24.18	5.39	11.71	6.18	12.03
Overlapping Windows [11]	15.26	28.69	6.72	11.63	13.68	26.40
Confidence Smoothing [34]	15.48	29.29	6.82	13.39	8.82	16.81
Frame-level Classification	15.61	32.88	1.84	4.49	2.79	7.27
Non-overlapping Windows [11]	17.07	34.07	7.40	13.44	13.85	27.17
BoCoNet	**18.06**	**35.25**	**7.87**	**14.01**	**15.40**	**28.29**

Average-AP. To assess the detection performance across various emotional categories, we introduce the metric Average-AP$_i$, which denotes the average detection results for emotion i at multiple IoU thresholds:

$$\text{Average-AP}_i = \frac{1}{J}\sum_{j=1}^{J} AP_{ij}. \tag{7}$$

5.2.3 Comparison Methods

We compare the proposed BoCoNet with several state-of-the-art methods of emotion recognition, expression spotting [27], [28], [29], [34], [35], and temporal emotion detection [11].

- **Frame-level Classification.** As a naive approach, the emotion category of each frame is predicted and adjacent frames with the same category are merged to generate emotion segments.
- **Feature Difference Analysis [35].** This micro-expression spotting method calculates the Chi-Squared distance of the LBP features of sequential frames within a specified interval. Then, threshold and peak detection are applied to locate the position of fast facial movements.
- **Optical Flow Magnitude [29].** As the first place in the Micro-Expression Grand Challenge 2022, this method locates the micro-expression and macro-expression events by calculating the optical flow magnitude in a sliding window.
- **Confidence Smoothing [34].** This method fashions the spotting task as a regression problem. After predicting and smoothing the confidence score at each moment, the standard threshold and peak detection techniques are utilized to spot the expression events in each video.
- **Non-overlapping Windows [11].** This method recognizes the emotion of each non-overlapping window and merges adjacent windows with the same emotion to get the final detection results.
- **Overlapping Windows [11].** As a temporal emotion detection method, windows of different lengths move in an overlapping fashion along the time axis, which generates multiple potential emotion segments. The completeness discrimination module is introduced to remove incomplete emotion segments.
- **Pyramid Sliding Window [28].** This method employs a predefined sliding window across videos of different temporal scales to generate a substantial number of candidate segments. A binary classifier and the NMS strategy are then utilized to eliminate background segments and remove overlapping segments, respectively.
- **Boundary Matching Network [27].** This approach leverages a complex boundary-matching mechanism to predict the confidence of facial expression segments and spot the expression segments.
- **VideoMAE-OF [53].** As the first place in the Micro-Expression Grand Challenge 2023, this method adopts the decision-level fusion strategy to integrate optical flow-based and VideoMAE [54]-based expression spotting results. To be consistent with other methods, we utilize the corresponding encoder instead of Video-MAE in the experiments.

5.3 Experimental Results

Table 4 shows the results of temporal emotion detection. It can be seen that for the Feature Difference Analysis [35] and Optical Flow Magnitude [29] methods, although traditional features (i.e., optical flow features and LBP features) have obtained satisfactory results on the expression spotting task in the laboratory environment, they lack robustness for in-the-wild temporal emotion detection tasks. The VideoMAE-OF method [53] improves performance in temporal emotion detection by incorporating the neural network-based feature with the optical flow feature. The Frame-level Classification method achieves a significant lead in TSAV features compared to Data2vec and Facepp features because TSAV contains temporal context information. Similarly, the Confidence Smoothing method [34], by integrating temporal dynamics within a window, acquires comparable results across all three features. Compared with the Pyramid Sliding Window method [28], the Overlapping Windows method [11] achieves a higher Average-mAP and Average-wAP, indicating that the completeness discrimination module is indispensable for detecting emotion segments with reasonable boundaries. The Boundary Matching Network [27], with its complex boundary-matching mechanism and hyperparameters, achieves limited performance in the task of temporal emotion detection. In contrast, the proposed BoCoNet, by predicting the boundaries of emotional segments and explic-

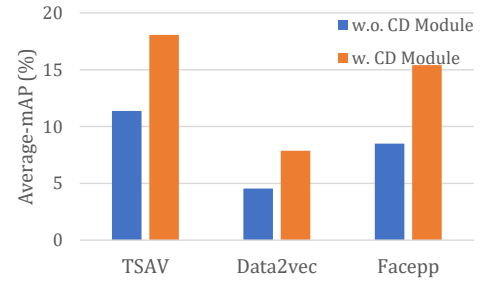
itly evaluating their completeness, obtains the best results across all metrics.

Moreover, for unimodal and multimodal representations, BoCoNet achieves the best results on TASV representation with an Average-mAP of 18.06%, and the worst results on Data2vec representation with an Average-mAP of 7.87%. This is because the supervised TASV and Facepp representations are specifically designed for expression recognition tasks. The self-supervised Data2Vec representation is trained on more general audio-visual data and contains less emotion-related information.

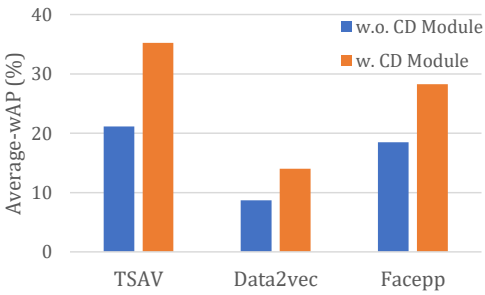
5.4 Ablation Study

5.4.1 Ablation on Module

To evaluate the effectiveness and necessity of the completeness discrimination (CD) module in BoCoNet, we further conduct module ablation experiments. Fig. 8 shows the experimental results on three features. It can be seen that in all features and metrics, removing the completeness discrimination module will bring significant performance degradation. The reason is that without the completeness discrimination module, BoCoNet would focus more on the most salient parts of the emotion segments and ignore the remaining parts, resulting in inaccurate boundary predictions. With the help of the completeness discrimination module, the model can effectively avoid these false positive samples and acquire better detection results.



(a) Average-mAP of module ablation experiments.

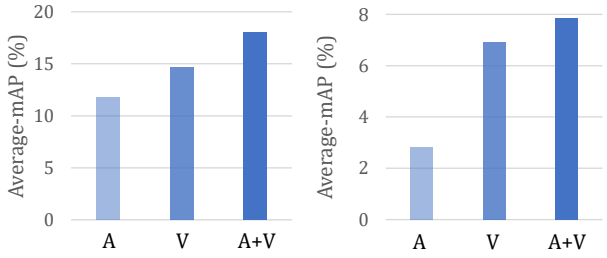


(b) Average-wAP of module ablation experiments.

Fig. 8. Detection results of module ablation experiments on three features.

5.4.2 Ablation on Modality

Fig. 9 shows the results of modality ablation experiments on two multimodal features. We can see that for both multimodal features, the visual modality (V) achieves better results compared to the auditory modality (A). This disparity may be attributed to the auditory modality containing more irrelevant information, such as background music or ambient noise, which does not contribute to emotion detection. In contrast, the visual modality directly conveys emotions through facial expressions, making it more effective for temporal emotion detection tasks. Furthermore, the fusion of audio and video modalities (A+V) can acquire better results. This improvement is expected as different modalities provide complementary cues that enhance the understanding of emotions, thereby boosting the overall performance of temporal emotion detection tasks.



(a) Results on TSAV. (b) Results on Data2vec.

Fig. 9. Detection results of modality ablation experiments on two multimodal features.

5.5 Further Analysis

5.5.1 Analysis of Detection Results by Emotion Category

To further analyze the detection performance on each emotion, we show the Average-AP of each emotion at multiple tIoU thresholds in Fig. 10.

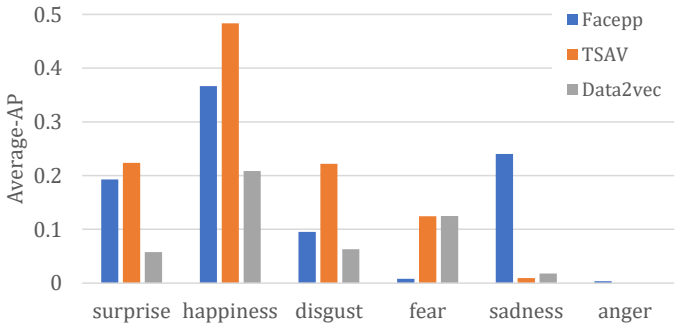


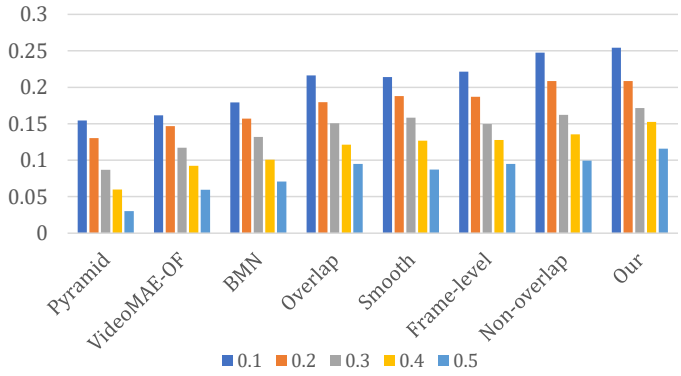
Fig. 10. Detection results on different emotion categories.

As the majority in the dataset, happiness, and surprise obtain better results on each feature. Sadness also achieves good performance on the Facepp feature, as well as disgust on the TSAV feature. Fear and anger achieve a relatively poor detection performance. Besides the limited amount of training data, the unremarkable expression of anger in the videos also increases the difficulty of detection. In conclusion, similar to the issue of imbalanced sample sizes, there is a significant variation in detection results across different emotion categories.

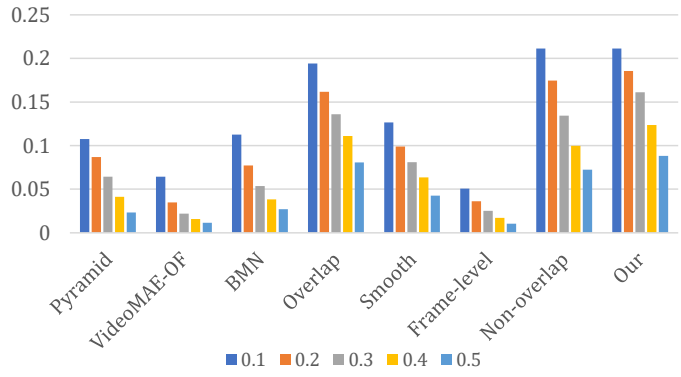
5.5.2 Analysis of Detection Results by tIoU Threshold

Fig. 11 shows the detection results at different tIoU thresholds on different features. We can observe that as the tIoU threshold increases, the detection performance decreases for all methods. The reason is that higher tIoU thresholds

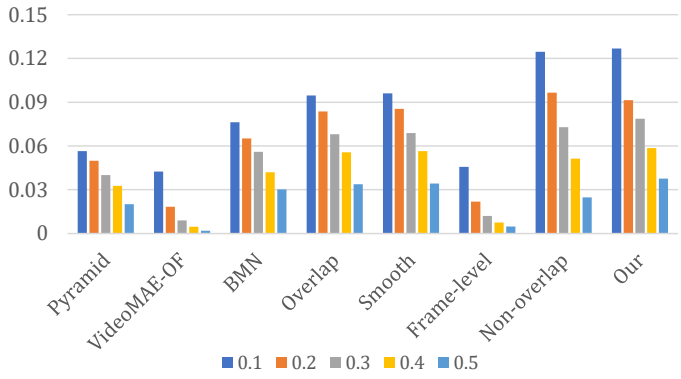
require more precise localization of the emotion segments, which is more challenging and prone to errors. Our method acquires the best or most competitive results on different thresholds and features, demonstrating its effectiveness and robustness.



(a) mAP at different tIoUs on TSAV.



(b) mAP at different tIoUs on Facepp.



(c) mAP at different tIoUs on Data2vec.

Fig. 11. Detection results at different tIoU thresholds.

5.5.3 Error Analysis

Analysing whether the errors come from the localization of starting/ending or the classification of emotions is important to further improve the detection performance. Table 5 shows the F1 scores for the classification results of starting, ending, and emotions on the test set, as well as the Average-mAP for temporal emotion detection. For emotion recognition, we clip the continuous videos in AVES following the boundaries of corresponding emotion segments, whose features are pooled and input into an MLP model pre-trained on the AVES training set.

TABLE 5
Error analysis of BoCoNet.

Encoder	Starting	Ending	Recognition	Detection
TSAV	**0.5998**	**0.5703**	**0.4605**	**0.1806**
Facepp	0.5253	0.5238	0.4409	0.1540
Data2vec	0.5452	0.5308	0.3980	0.0787

We can see that the TASV feature obtains the best results on both boundary classification and emotion classification, and finally achieves a 2.66% improvement on temporal emotion detection compared to the Facepp feature. This indicates the performances of boundary localization and emotion classification influence the final temporal emotion detection result. Classification of the starting boundaries achieves better results than the ending boundaries, which is consistent with the psychological research that emotions have a clear onset and a somewhat fuzzy offset [55].

5.5.4 Qualitative analysis

Fig. 12 illustrates two successfully detected emotion segments and two unsuccessfully detected emotion segments. For the two successfully detected emotion segments, the tIoU between the prediction and ground truth is 0.829 and 0.818, respectively, which is higher than the threshold of 0.5, and the emotion categories are correctly classified. For the first unsuccessfully detected emotion segment, the emotion category of this segment is misclassified as *sadness*, which is inconsistent with the ground-truth *surprise*. For the second unsuccessfully detected emotion segment, the tIoU between the prediction and ground truth is 0.377, which is below the threshold of 0.5. Therefore, these two emotion segments are considered as undetected segments.

6 Discussion of Ethics and Conclusion

6.1 Discussion of Ethics

The AVES dataset, similar to the other in-the-wild emotion datasets [18], [23], [24], is collected from online video sites and contains personal information such as facial images or speech. Thus, there is a potential risk of privacy violation. To address these concerns, we will release the pre-extracted unimodal and multimodal features to the research community instead of the raw audio-visual data.

The proposed dataset can have a positive impact on society as it can provide data and a test bed for detecting and localizing emotion segments from continuous signals, which can be applied to potential mental health or human-computer applications [3], [56]. However, it is important to acknowledge the possible negative social impact due to privacy violation risks. We have recognized the significance of ethical considerations and will continue to monitor and enhance our privacy protection measures.

6.2 Conclusion

The challenge of emotion understanding in real-world scenarios lies in the fact that emotions are constantly changing, and the corresponding data is continuous and untrimmed. In this paper, we propose an audio-visual emotion stream

Fig. 12. Visualization of two detected segments and two undetected segments.

(AVES) dataset for the challenging task of temporal emotion detection. Unlike existing datasets that focus on utterance-level emotion recognition or frame-level facial expression recognition, AVES provides a benchmark for locating the boundaries of emotion segments and predicting the emotion category from continuous signals. This dataset contains in-the-wild videos with rich emotional changes and has been reliably annotated with the time boundaries and categories of each emotion segment. Furthermore, we propose a boundary combination network (BoCoNet) to detect emotion segments with more flexible boundaries and durations. Experimental results show that the proposed BoCoNet achieves the best performance on several multimodal and unimodal features.

Moreover, in the course of conducting this research, some critical challenges, which we believe are important to address in future research on temporal emotion detection, are identified:

1) **Inter-segment relationship modeling**. This study explores the effect of short-term context on temporal emotion detection. However, there might be potential relations between different emotion segments in a continuous signal [57], [58]. Therefore, exploring the relationship between different emotion segments could help to localize and recognize the emotion segments.

2) **One-stage temporal emotion detection**. Existing temporal emotion methods typically involve first localizing emotional segments and then recognizing emotional categories. Inspired by related work on object detection [59], [60], developing a one-stage temporal emotion detection framework for simultaneous localization and recognition is a promising research direction. Compared to two-stage detection frameworks, the one-stage detection framework offers a faster processing speed.

3) **Multimodal representation learning**. Experimental results show that even with the same detection framework, the detection performance can vary across different representation sequences. This motivates us to explore more effective multimodal representation learning and fusion methods for temporal emotion detection [61], [62], [63].

Overall, temporal emotion detection, as a novel but challenging task, has great potential for further exploration.

ACKNOWLEDGMENTS

This paper is supported by the National Natural Science Foundation of China (grant 62236006), and the Key Research and Development Program of Shaanxi (No. 2022ZDLGY06-03).

REFERENCES

[1] R. W. Picard, *Affective computing*. MIT press, 2000.
[2] F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt, S. Alisamir, S. Amiriparian, E.-M. Messner *et al.*, "Avec 2019 workshop and challenge: State-of-mind, depression with ai, and cross-cultural affect recognition," in *Proceedings of the 9th International Workshop on Audio/Visual Emotion Challenge, AVEC*, vol. 19, 2019.
[3] Y. Li, T. Yang, L. Yang, X. Xia, D. Jiang, and H. Sahli, "A multimodal framework for state of mind assessment with sentiment pre-classification," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, 2019, pp. 13–18.
[4] A. M. Rahmani, J. Lai, S. Jafarlou, I. Azimi, A. Yunusova, A. Rivera, S. Labbaf, A. Anzanpour, N. Dutt, R. Jain *et al.*, "Personal mental health navigator: Harnessing the power of data, personal models, and health cybernetics to promote psychological well-being," *Frontiers in Digital Health*, vol. 4, p. 933587, 2022.
[5] X. Xia and D. Jiang, "Hit-mst: Dynamic facial expression recognition with hierarchical transformers and multi-scale spatiotemporal aggregation," *Information Sciences*, p. 119301, 2023.
[6] R. Zhao, T. Liu, Z. Huang, D. P. Lun, and K.-M. Lam, "Spatial-temporal graphs plus transformers for geometry-guided facial expression recognition," *IEEE Transactions on Affective Computing*, 2022.
[7] Z. Lian, B. Liu, and J. Tao, "Smin: Semi-supervised multi-modal interaction network for conversational emotion recognition," *IEEE Transactions on Affective Computing*, 2022.
[8] P. Antoniadis, I. Pikoulis, P. P. Filntisis, and P. Maragos, "An audiovisual and contextual approach for categorical and continuous emotion recognition in-the-wild," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2021, pp. 3645–3651.
[9] J.-H. Kim, N. Kim, and C. S. Won, "Facial expression recognition with swin transformer," *arXiv preprint arXiv:2203.13472*, 2022.
[10] A. Psaroudakis and D. Kollias, "Mixaugment & mixup: Augmentation methods for facial expression recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2022, pp. 2367–2375.

[11] Y. Li, X. Xia, D. Jiang, H. Sahli, and R. Jain, "Memos: A multimodal emotion stream database for temporal spontaneous emotional state detection," in *Companion Publication of the 2020 International Conference on Multimodal Interaction*, 2020, pp. 370–378.

[12] Y. Wang and L. Guan, "Recognizing human emotional state from audiovisual signals," *IEEE transactions on multimedia*, vol. 10, no. 5, pp. 936–946, 2008.

[13] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The enterface'05 audio-visual emotion database," in *22nd International Conference on Data Engineering Workshops (ICDEW'06)*. IEEE, 2006, pp. 8–8.

[14] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[15] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," *arXiv preprint arXiv:1810.02508*, 2018.

[16] X. Jiang, Y. Zong, W. Zheng, C. Tang, W. Xia, C. Lu, and J. Liu, "Dfew: A large-scale database for recognizing dynamic facial expressions in the wild," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 2881–2889.

[17] Y. Wang, Y. Sun, Y. Huang, Z. Liu, S. Gao, W. Zhang, W. Ge, and W. Zhang, "Ferv39k: a large-scale multi-scene dataset for facial expression recognition in videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 922–20 931.

[18] D. Kollias and S. Zafeiriou, "Aff-wild2: Extending the aff-wild database for affect recognition," *arXiv preprint arXiv:1811.07770*, 2018.

[19] J. Yang, Q. Huang, T. Ding, D. Lischinski, D. Cohen-Or, and H. Huang, "Emoset: A large-scale visual emotion dataset with rich attributes," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 20 383–20 394.

[20] F. Qu, S.-J. Wang, W.-J. Yan, H. Li, S. Wu, and X. Fu, "Cas(me)2: a database for spontaneous macro-expression and micro-expression spotting and recognition," *IEEE Transactions on Affective Computing*, vol. 9, no. 4, pp. 424–436, 2017.

[21] J. Li, Z. Dong, S. Lu, S.-J. Wang, W.-J. Yan, Y. Ma, Y. Liu, C. Huang, and X. Fu, "Cas (me) 3: A third generation facial spontaneous micro-expression database with depth information and high ecological validity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[22] Z. Zhang and J. Yang, "Temporal sentiment localization: Listen and look in untrimmed videos," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 199–208.

[23] A. Zadeh and P. Pu, "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, 2018.

[24] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," *arXiv preprint arXiv:1606.06259*, 2016.

[25] B. Xu, Y. Fu, Y.-G. Jiang, B. Li, and L. Sigal, "Heterogeneous knowledge transfer in video emotion recognition, attribution and summarization," *IEEE Transactions on Affective Computing*, vol. 9, no. 2, pp. 255–270, 2016.

[26] Y.-G. Jiang, B. Xu, and X. Xue, "Predicting emotions in user-generated videos," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 28, no. 1, 2014.

[27] X. Xu, Y. Zong, W. Zheng, Y. Li, C. Tang, X. Jiang, and H. Jiang, "Sdfe-lv: A large-scale, multi-source, and unconstrained database for spotting dynamic facial expressions in long videos," *arXiv preprint arXiv:2209.08445*, 2022.

[28] T.-K. Tran, X. Hong, and G. Zhao, "Sliding window based micro-expression spotting: a benchmark," in *international conference on advanced concepts for intelligent vision systems*. Springer, 2017, pp. 542–553.

[29] J. Yu, Z. Cai, Z. Liu, G. Xie, and P. He, "Facial expression spotting based on optical flow features," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 7205–7209.

[30] Y. Zhao, X. Tong, Z. Zhu, J. Sheng, L. Dai, L. Xu, X. Xia, Y. Jiang, and J. Li, "Rethinking optical flow methods for micro-expression spotting," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 7175–7179.

[31] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Collecting large, richly annotated facial-expression databases from movies," *IEEE multimedia*, vol. 19, no. 03, pp. 34–41, 2012.

[32] T.-K. Tran, Q.-N. Vo, X. Hong, X. Li, and G. Zhao, "Micro-expression spotting: A new benchmark," *Neurocomputing*, vol. 443, pp. 356–368, 2021.

[33] T.-K. Tran, Q.-N. Vo, X. Hong, and G. Zhao, "Dense prediction for micro-expression spotting based on deep sequence model," *Electronic Imaging*, vol. 2019, no. 8, pp. 401–1, 2019.

[34] G.-B. Liong, J. See, and L.-K. Wong, "Shallow optical flow three-stream cnn for macro-and micro-expression spotting from long videos," in *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021, pp. 2643–2647.

[35] X. Li, X. Hong, A. Moilanen, X. Huang, T. Pfister, G. Zhao, and M. Pietikäinen, "Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods," *IEEE transactions on affective computing*, vol. 9, no. 4, pp. 563–577, 2017.

[36] J. Li, M. H. Yap, W.-H. Cheng, J. See, X. Hong, X. Li, S.-J. Wang, A. K. Davison, Y. Li, and Z. Dong, "Megc2022: Acm multimedia 2022 micro-expression grand challenge," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 7170–7174.

[37] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikäinen, "A spontaneous micro-expression database: Inducement, collection and baseline," in *2013 10th IEEE International Conference and Workshops on Automatic face and gesture recognition (fg)*. IEEE, 2013, pp. 1–6.

[38] W.-J. Yan, Q. Wu, Y.-J. Liu, S.-J. Wang, and X. Fu, "Casme database: A dataset of spontaneous micro-expressions collected from neutralized faces," in *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*. IEEE, 2013, pp. 1–7.

[39] R. Plutchik, "A general psychoevolutionary theory of emotion," in *Theories of emotion*. Elsevier, 1980, pp. 3–33.

[40] A. Dhall, R. Goecke, J. Joshi, J. Hoey, and T. Gedeon, "Emotiw 2016: Video and group-level emotion recognition challenges," in *Proceedings of the 18th ACM international conference on multimodal interaction*, 2016, pp. 427–432.

[41] F. Kuhnke, L. Rumberg, and J. Ostermann, "Two-stream aural-visual affect analysis in the wild," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. IEEE, 2020, pp. 600–605.

[42] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-nms–improving object detection with one line of code," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5561–5569.

[43] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[44] W.-J. Yan, Q. Wu, J. Liang, Y.-H. Chen, and X. Fu, "How fast are the leaked facial expressions: The duration of micro-expressions," *Journal of Nonverbal Behavior*, vol. 37, no. 4, pp. 217–230, 2013.

[45] D. Matsumoto and H. S. Hwang, "Evidence for training the ability to read microexpressions of emotion," *Motivation and emotion*, vol. 35, no. 2, pp. 181–191, 2011.

[46] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "Data2vec: A general framework for self-supervised learning in speech, vision and language," in *International Conference on Machine Learning*. PMLR, 2022, pp. 1298–1312.

[47] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[48] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[49] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[50] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[51] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[52] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.
[53] K. Xu, K. Chen, L. Sun, Z. Lian, B. Liu, G. Chen, H. Sun, M. Xu, and J. Tao, "Integrating videomae based model and optical flow for micro-and macro-expression spotting," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 9576–9580.
[54] Z. Tong, Y. Song, J. Wang, and L. Wang, "Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training," *Advances in neural information processing systems*, vol. 35, pp. 10 078–10 093, 2022.
[55] K. R. Scherer, "Emotions are emergent processes: they require a dynamic computational architecture," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 364, no. 1535, pp. 3459–3474, 2009.
[56] Y. Li, L. Zhang, X. Lan, and D. Jiang, "Towards adaptable graph representation learning: An adaptive multi-graph contrastive transformer," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 6063–6071.
[57] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, "Dialoguernn: An attentive rnn for emotion detection in conversations," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 6818–6825.
[58] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh, "Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 154–164.
[59] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer, 2016, pp. 21–37.
[60] D. Li, A. Wu, Y. Wang, and Y. Han, "Prompt-driven dynamic object-centric learning for single domain generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 17 606–17 615.
[61] Y. Li, X. Lan, H. Chen, K. Lu, and D. Jiang, "Multimodal pear chain-of-thought reasoning for multimodal sentiment analysis," *ACM Transactions on Multimedia Computing, Communications and Applications*, 2024.
[62] L. Xiao, X. Yang, F. Peng, M. Yan, Y. Wang, and C. Xu, "Clip-vg: Self-paced curriculum adapting of clip for visual grounding," *IEEE Transactions on Multimedia*, 2023.
[63] L. Xiao, X. Yang, F. Peng, Y. Wang, and C. Xu, "Hivg: Hierarchical multimodal fine-grained modulation for visual grounding," *arXiv preprint arXiv:2404.13400*, 2024.