SIR-HCL: Semantic-Inconsistency Reasoning and Hybrid Contrastive Learning for Efficient Cross-Emotion Anomaly Detection

Xin Liu, *Senior Member, IEEE*, Qiyan Chen, Yiu-ming Cheung, *Fellow, IEEE*, and Shu-Juan Peng, *Senior Member, IEEE*

***Abstract*—Cross-emotion anomaly detection is an emerging and challenging research topic in cognitive analysis field, which aims at identifying the abnormal emotion pair whose semantic patterns are inconsistent across different emotional modalities. To the best of our knowledge, this topic has yet to be well studied, which could potentially benefit lots of valuable cognitive applications such as autistic children diagnosis and criminal deception detection. To this end, this article proposes an efficient cross-emotion anomaly detection approach via semantic-inconsistency reasoning and hybrid contrastive learning (SIR-HCL), which is the first attempt to detect the anomalous emotional pairs across the audio–visual emotions. First, the proposed framework utilizes dual-branch network to obtain the deep emotional features in each modality, and then employs the shared residual block to derive the semantically compatible features. Subsequently, an efficient hybrid contrastive learning approach is designed to enlarge the semantic-inconsistency among abnormal emotional pair with different affective classes, while enhancing the semantic-consistency and increasing the feature correlation between normal emotional pair from the same affective class. At the same time, an efficient bidirectional learning scheme is employed to significantly improve the data utilization and a two-component Beta Mixture Model is adaptively utilized to reason the anomalous emotion pairs. Extensive experiments evaluated on two benchmark datasets show that the proposed SIR-HCL method can well detect the anomalous emotional pairs across audio-visual emotional data, and brings substantial improvements over the state-of-the-art competing methods.**

***Index Terms*—Audio–visual emotion, beta mixture model, cross-emotion anomaly detection, hybrid contrastive learning, semantic-inconsistency reasoning.**

I. Introduction

EMOTION recognition is one of the most attractive interdisciplinary research topics in artificial intelligence [1], which has drawn much attention recently and led to the advancement of a wide range of applications, such as sentiment analysis, psychological evaluation, the assessment of fatigue and depression. Many cognitive studies lend sufficient credence to the hypothesis that the perception of human emotion plays a vital role in their everyday lives. To be specific, anomaly detection in emotional data refers to identifying the human's abnormal emotional patterns that are significantly different from other numerous normal emotional patterns, which is an important sentiment analysis technique due to the fact that the anomalous emotions often provide significant and critical information to the evaluation of psychological counseling, autism diagnosis, and healthcare treatments [2]. For instance, many studies in cognitive science have shown that anomaly detection in facial emotions is of crucial importance to the evaluation of depression, while anomalous pattern identification in voice-conversation plays an important role in diagnosing and screening autistic children.

In the literature, most existing abnormal emotion detection methods predominately focus on examining the emotional data from a single source, e.g., facial data, voice data, or social media data [3]. In recent years, there has been a growing interest in the research of multimodal emotion analysis, due to its potential in providing rich information and robustness to sensor noise [4]. Under such circumstances, anomaly detection from multimodal emotional data is highly desirable in many applications such as disease monitoring and abnormal behavior analysis. For instance, anomaly detection in multimodal

Received 19 August 2024; revised 15 February 2025; accepted 5 March 2025. Date of publication 12 March 2025; date of current version 15 October 2025. This work was supported in part by the National Science Foundation of China under Grant 62476103, in part by the Natural Science Foundation of Xiamen City under Grant 3502Z202473043, in part by the National Science Foundation of Fujian Province under Grant 2024J01096 and Grant 2022J01316, in part by the NSFC/Research Grants Council (RGC) Joint Research Scheme under Grant N_HKBU214/21, in part by the RGC Senior Research Fellow Scheme under Grant SRFS2324-2S02, and in part by the General Research Fund of RGC under Grant 12201321 and Grant 12202622. Recommended for acceptance by Associate Editor Guoqi Li. *(Corresponding authors: Yiu-ming Cheung; Shu-Juan Peng.)*

Xin Liu is with the Department of Computer Science, Huaqiao University, Xiamen 361021, China, and also with the Department of Computer Science, Hong Kong Baptist University, Hong Kong SAR, China (e-mail: xliu@hqu.edu.cn).

Qiyan Chen is with Xiamen Key Laboratory of Computer Vision and Pattern Recognition, Xiamen 361021, China, and also with Fujian Key Laboratory of Big Data Intelligence and Security, Huaqiao University, Xiamen 361021, China (e-mail: qychen@hqu.edu.cn).

Yiu-ming Cheung is with the Department of Computer Science, Hong Kong Baptist University, Hong Kong SAR, China (e-mail: ymc@comp.hkbu.edu.hk).

Shu-Juan Peng is with the Department of Artificial Intelligence, Huaqiao University, Xiamen 361021, China, and also with the Key Laboratory of Computer Vision and Machine Learning (Huaqiao University), Fujian Province University, Xiamen 361021, China (e-mail: pshujuan@hqu.edu.cn).

Digital Object Identifier 10.1109/TCDS.2025.3550645

conversation data focuses on identifying abnormal sentiment patterns and their temporal dynamics [5], which can be well utilized to perform psychological evaluation on the speaker. Moreover, the analysis of multi-modal emotional data collected by innovative imaging sensors and user interactions can provide automatic remote monitoring of Parkinson's patients, aiding in early health-related event detection. Besides, recent cognitive studies have shown substantial evidence that autistic children struggle with cross-modal integration during expressive tasks [6]. Nevertheless, the subtle movements in facial dynamics may not be easily captured by manual visual inspection alone. As such, traditional unimodal abnormal emotion detectors cannot discover multimodal emotional anomalies.

In fact, it is found that some of the multimodal emotions are often not anomalous when they are viewed separately in each individual modality, but which contain inconsistent responses when multiple modalities are jointly considered. Audio–visual emotion may probably be the most natural multi-modal expression to achieve emotion analysis, favored for its unique advantages including ease of use and is less invasive to humans [7], [8]. In the literature, most recent audio–visual emotion analysis methods primarily concentrate on fusing the heterogeneous features extracted from facial and vocal modalities, with limited exploration of semantic inconsistencies across different emotional modalities. Although recent multiview anomaly detection algorithms have been designed to detect the anomalous samples that have abnormal behaviors in each view or have inconsistent behaviors across different views, they are not directly applicable to dynamic and heterogeneous audio–visual emotional data. Besides, there has been little discussion about semantic-inconsistency analysis across different emotional modalities.

In this article, we focus on a relatively new topic in the abnormal emotion detection field, i.e., cross-emotion anomaly detection, particularly for audio–visual emotional data. It aims at identifying the abnormal emotional data pair whose affective patterns are inconsistent across different emotional modalities, which may benefit lots of valuable cognitive applications such as psychological disease diagnosis and abnormal behavior monitoring. For instance, suspects may attempt to hide their emotions in criminal or judicial cases, whose emotions might have incongruity between facial microexpression and speech emotion. Under such circumstances, abnormal emotion detection across different emotional modalities is beneficial to assist potential deception detection. Besides, recent cognitive studies have shown substantial evidence that autistic children often produce emotional sentences with weak cross-modal consistency across speech and facial expressions [6]. As shown in Fig. 1, the emotions conveyed through their facial expressions could be positive, while the relevant emotions surveyed by their speeches are negative. Evidently, developing a computational approach to detecting the cross-modal emotional inconsistency, if any, is capable of providing a new promising way for the early screening of autistic children.

To the best of our knowledge, cross-emotion anomaly detection across audio-visual modalities has yet to be well studied and there are still three main challenges. 1) Weak emotional representation: the multimodal emotions acquired from different modalities may cover different dynamic representations, and it is generally difficult to extract the most relevant, comprehensive and discriminative expression from each modality [9]. 2) Modality heterogeneity: audio and visual emotions are captured by different sensors, and there exists a huge modality gap between audio and visual emotion representations. 3) Complex semantic correlation: most existing audio–visual emotion analysis methods often fuse the information from various modalities to learn a richer multimodal representation, which inherently ignores the semantic relevance and difference between different emotional modalities. Therefore, it is still a nontrivial task to perform efficient cross-emotion anomaly detection from a practical viewpoint.

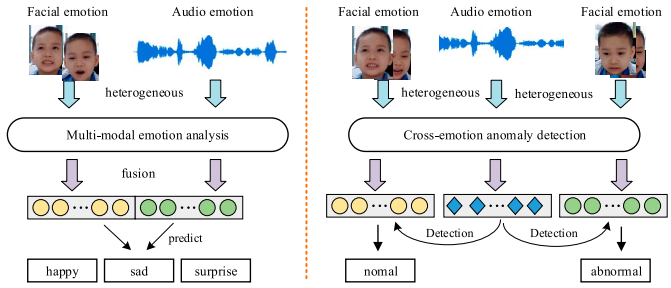

Fig. 1. Illustration of the difference between multimodal emotion analysis and cross-emotion anomaly detection.

In this article, we propose an efficient cross-emotion anomaly detection framework via semantic-inconsistency reasoning and hybrid contrastive learning (SIR-HCL), which is the first attempt to detect the anomalous emotions across the audio–visual data. The proposed framework can well reason the semantic-inconsistency among the high-level audio–visual representations of all semantically irrelevant emotions, while enhancing the semantic-consistency between the semantically relevant ones. The main contributions are summarized as follows.

1) A novel cross-emotion anomaly detection framework is explicitly designed to identify the abnormal emotions across audio–visual data. To the best of our knowledge, this work is the first attempt to detect the anomalous audio–visual emotions which do not behave consistently across different modalities.
2) An efficient hybrid contrastive learning method is presented to simultaneously preserve intra-/cross-modal interactions and hard-sample relationships, which can well learn the discriminative cross-emotion embeddings by contrastive learning in a hybrid manner.
3) The bidirectional learning scheme is developed to significantly improve the data utilization, while a two-component Beta Mixture Model is well designed to reason about semantic-inconsistency and the semantic-consistency across different emotional modalities.
4) Extensive experiments verify the superiorities of the proposed framework and show its competitive abnormal emotion detection performances.

The remainder part of this article is structured as follows: Section II provides a brief overview of related works. In Section III, we elaborate the architecture and implementation details of the proposed framework, and Section IV presents the experimental results and extensive comparisons. Finally, we draw a conclusion in Section V.

II. Related Works

Cross-emotion anomaly detection across audio–visual data is a relatively new research topic in cognitive science, and this section mainly surveys the most relevant abnormal emotion detection or multimodal anomaly detection works.

A. Abnormal Emotion Detection

Abnormal emotion detection primarily refers to identify the abnormal sentiments, opinions, or attitudes from numerous normal patterns. Often, the abnormal emotion may be hidden in a facial expression, voice conversation, or a paragraph of text to reflect sudden changes in sentiment. Along this line, Germine et al. [10] employed the functional magnetic resonance imaging to investigate abnormal neural activity in emotional face processing, while Clavel et al. [11] developed a fear-type emotion recognition system to detect abnormal situations for surveillance applications. Later, Sun et al. [5] utilized a hybrid model that combines the convolutional neural network, long short-term memory network and Markov chain Monte Carlo (MCMC) methods to identify conversation anomaly.

The detection of abnormal emotions using physiological signals also brings significant benefits to the field of digital healthcare and human–computer interaction. Along this way, Gannouni et al. [12] utilize dthe electrocardiogram signals to detect abnormal emotions and therefore bring significant benefits to the field of digital healthcare. Later, Zhu et al. [13] utilized a low-cost wearable sensor to collect electrocardiogram signals and present an unsupervised abnormal emotion detection method. Note that, these methods mainly focus on detecting the unusual samples from a single emotional data. With the popularity of different sensors, multimodal emotion learning has gained increasing attention due to the availability of diverse information sources [14]. For instance, Alvarez et al. [15] gathered emotional signals from multiple sources to detect abnormal behaviors, offering critical insights for early health-related event prevention. Nevertheless, this approach focuses on identifying anomalies when emotional signals from one modality are missing or corrupted, which essentially neglects the affective relationships between different emotional modalities and therefore cannot identify the potential semantic inconsistencies across different emotional modalities.

B. Multimodal Anomaly Detection

Multimodal data analysis mainly aggregates both independent and complementary information to provide comprehensive representations, and such a technique has drawn much interest in multimodal emotion analysis [16], [17]. For instance, Chen et al. [17] proposed a hybrid fusion based on information relevance (HFIR) for multimodal sentiment analysis, which unifies two separate multimodal networks to mine the complementary and correlated information among different modalities. Notably, this approach fuses the features of different emotions to perform sentiment analysis, is incapable of discovering the inherent anomalous behaviors across different emotional modalities. Specifically, multimodal abnormal emotion detectors mainly aim to identify the possible anomalies from completely heterogeneous emotional modalities, such as visual, audio, and text data. Along this line, Dawel et al. [18] utilized the meta-analyses to detect the evidence of pervasive impairments across facial and vocal modalities, with significant deficits evident for several emotions (i.e., not only fear and sadness) in both adults and children/adolescents. Note that, this approach just employs a low-level fusion method to detect the significant impairments between the facial and vocal expressions, which cannot reveal fine-grained abnormal expressions in the captured multimodal emotions. In recent years, multimodal deep-learning based anomaly detection algorithms have become increasingly popular, and some works have attempted to cast the anomaly detection problem as a one-class classification problem or as the detection of out-of-distribution samples. Along this way, Jiang et al. [19] utilize the deep networks to interact visual and auditory signals, and jointly create a sense of emotional atmosphere within the scene. Accordingly, they further build an audio–visual modality fusion model to recognize the abnormal emotion. It is noted that such a method mainly fuses multiple information to capture the abnormal information, which neglects the semantic relationship and interactions between different emotional modalities. Therefore, this approach is explicitly incapable of discovering the abnormal multimodal emotions that have inconsistent behaviors across different emotional modalities.

Recently, Li et al. [20] have innovated the concept of cross-modal anomaly detection (CMAD), which aims to identify the inconsistent patterns or behaviors of instances across different modalities. Specifically, this approach first trains multimodal deep neural networks to extract features from different modalities, and then utilizes a predefined threshold to detect the potential cross-modal anomalies. Note that, this method is tailored to detect multimodal data with substantial semantic inconsistencies, which is incapable of identifying dynamic cross-modal anomalies, including inconsistent behaviors across facial expressions or acoustic patterns. Therefore, there is still a lack of efficient models to achieve cross-emotion anomaly detection from a practical viewpoint.

III. Methodology

Cross-emotion anomaly detection is a relatively new topic in multimodal emotion analysis field. Without loss of generality, the proposed framework mainly focuses on abnormal emotion detection across audio–visual data pairs, and this section first clarifies the relevant notation and formal definition of cross-emotion anomaly detection. Then, the proposed network architecture, hybrid contrastive learning, and semantic-inconsistency mining scheme are introduced in tandem. Finally, the reasoning

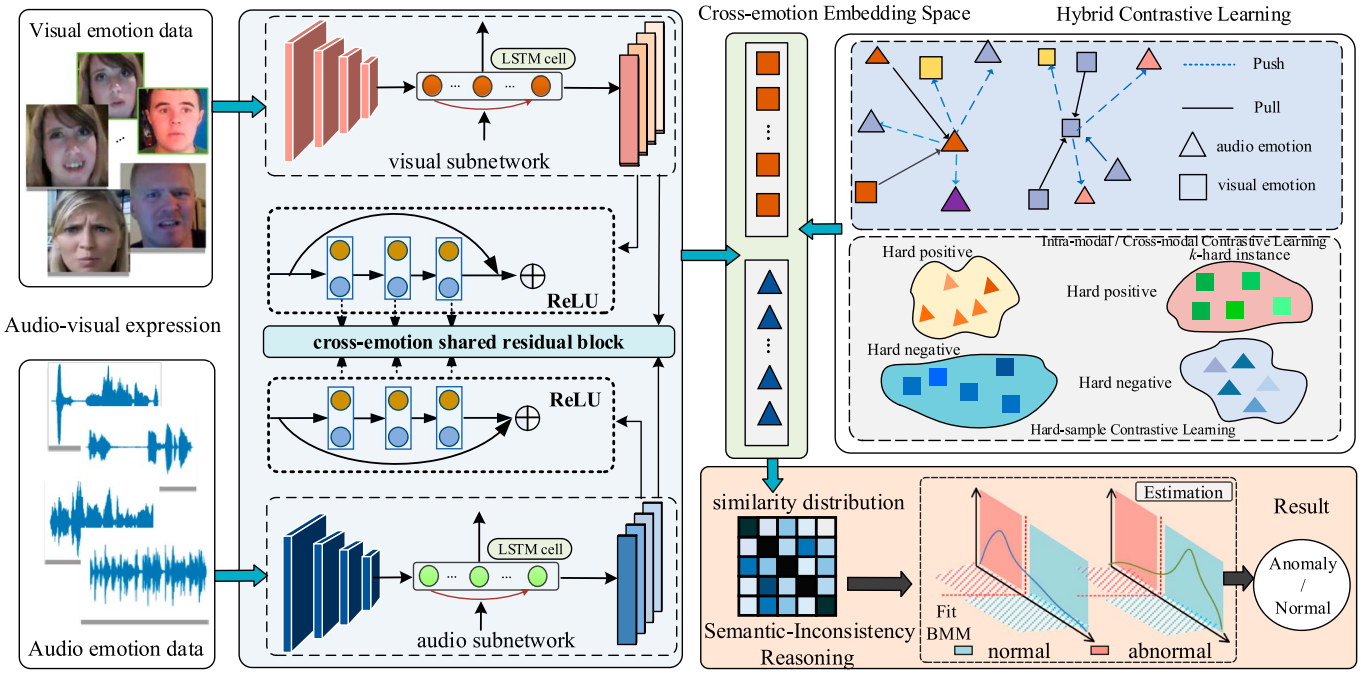

Fig. 2. Schematic architecture of the proposed cross-emotion anomaly detection framework.

of anomalous emotional data pair and its optimization process are explicitly provided.

A. Notation and Problem Formulation

Suppose that we have an emotional multimodal dataset that consists of visual emotional data $\mathbf{X}^v = \{\mathbf{x}_i^v\}_{i=1}^N$ and audio emotional data $\mathbf{X}^a = \{\mathbf{x}_i^a\}_{i=1}^N$, where $\mathbf{x}_i^v \in \mathbb{R}^{l \times d_v}$ and $\mathbf{x}_i^a \in \mathbb{R}^{l \times d_a}$, N represent the total data number within these two modalities, l is the emotional sequence length, d_a and d_v are respectively the raw embedding dimension of visual and audio modalities. Given an emotional data pair $\{\mathbf{x}_i^v, \mathbf{x}_i^a\}$, the primary task of cross-emotion anomaly detection is to determine whether the emotional data pair $\mathbf{x}_i^v$ and $\mathbf{x}_i^a$ have inconsistent emotional behavior or not.

Note that, the training of a single-modal classifier on visual data or audio data is able to predict the affective labels of individual emotion instance, and these predicted affective labels can be intuitively utilized to evaluate the semantic-inconsistency between emotional data pairs. It is noteworthy that such a training approach inherently necessitates the explicit definition and prior prediction of affective labels for each emotion instance. However, if the emotional instances are inaccurately classified into incorrect categories, this baseline model will make a wrong prediction and fail to detect semantic-inconsistency across different emotional modalities. To tackle this problem, as shown in Fig. 2, we present an efficient cross-emotion anomaly detection framework via semantic-inconsistency reasoning and hybrid contrastive learning (SIR-HCL). Specifically, we formulate the cross-emotion anomaly detection as a binary classification problem, and utilize the semantic-consistency labels of emotional data pairs to measure the detection results, thereby bypassing the prediction of affective categories for each single emotion instances. To the best of our knowledge, the proposed SIR-HCL framework is the first attempt to detect cross-modal emotional anomalies within heterogeneous emotional expressions.

B. Dual-Branch Network Architecture

The proposed SIR-HCL model aims to identify semantic-inconsistency among high-level representations of irrelevant expressions while enhancing semantic-consistency within relevant ones. Recent advances in multimodal deep neural networks have enabled effective learning of compatible features across modalities [21]. Specifically, we first utilize a dual-branch network architecture to obtain the deep emotional feature representations in each modality, and then employ the shared residual block to derive the semantically compatible features across heterogeneous emotions.

Feature Encoding Subnetworks: The inputs to the visual subnetwork and audio subnetwork are respectively the facial image sequence and acoustic frames in an emotional utterance. Specifically, the visual subnetwork $\mathcal{F}_v(\cdot)$ and audio subnetwork $\mathcal{F}_a(\cdot)$ are both implemented using a single-directional long short-term memory (LSTM), and the end-state hidden representations followed by a fully connected layer are selected as the outputs. To produce an efficient visual emotion embedding that has the same size as the acoustic emotion embedding, we set the size of the last network layer to be the same for each subnetwork. Consequently, the high-level visual emotional embedding $\mathbf{v}_i$ and acoustic emotional embedding $\mathbf{a}_i$ of the ith data pair can be obtained by

$$\mathbf{v}_i = \mathcal{F}_v(\mathbf{x}_i^v, \theta_v^{\text{lstm}}), \mathbf{a}_i = \mathcal{F}_a(\mathbf{x}_i^a, \theta_a^{\text{lstm}}) \quad (1)$$

where θ_v^{lstm} and θ_a^{lstm} respectively represent network parameters of visual subnetwork and audio subnetwork.

Shared Residual Block: The heterogeneous audio and visual feature representations often exhibit significantly different dynamic characteristics, and previous multi-modal emotion analysis works often learn modality-invariant and modality-specific features for predicting the affective states. Although audio–visual emotional embeddings may be semantically relevant at the utterance level, there still exists a modality gap across two modalities. To alleviate this concern, we employ a shared residual block to learn the semantically relevant representation of heterogeneous embeddings and bridge their semantic gap between different emotional modalities

$$\mathbf{v}_i^r = \sigma\left(\mathbf{v}_i + \eta \cdot \texttt{FC}(\mathbf{v}_i)\right), \mathbf{a}_i^r = \sigma\left(\mathbf{a}_i + \eta \cdot \texttt{FC}(\mathbf{a}_i)\right) \tag{2}$$

where σ is the nonlinear activation function such as ReLU, η represents a learnable scale factor, and FC denotes the fully connected layer. Consequently, the shared residual block connects the output features in jumps, which can be well utilized to preserve the original features as well as mitigating the elimination of network gradients. Meanwhile, the shared weighting parameters through fully connected layers are beneficial to achieve cross-emotion compatible feature learning, and such a learnable residual block structure can be optimized by the shared embedding loss $\mathcal{L}_{\text{ice}}$

$$\mathcal{L}_{\text{share}} = \mathcal{L}_{\text{ce}}(\mathbf{v}_i^r, y_i^v) + \mathcal{L}_{\text{ce}}(\mathbf{a}_i^r, y_i^a) \tag{3}$$

where y_i^v and y_i^a are respectively the ground truth of affective labels for the ith emotional pair, $\mathcal{L}_{\text{ce}}$ is the symbolic notation of standard cross-entropy loss.

C. Hybrid Contrastive Learning Module

Contrastive learning has emerged as a highly effective approach for representation learning, which allows the model to contrast the positive samples against a set of negative samples [22]. For instance, Kim et al. [23] presented a self-supervised contrastive learning framework to train a robust neural network without labeled data, while Yang et al. [24] combined the spike-based self-supervised learning and contrastive learning to train the spiking neural network. Besides, Wu et al. [25] provided a comprehensive review of existing self-supervised contrastive learning techniques for graph data. Notably, these contrastive learning works are generally designed to enhance representation learning in single modality. The proposed framework aims to enlarge the semantic-inconsistency between emotional data pair with different affective classes, while enhancing the semantic-consistency between emotional data from the same affective class. To this end, we present a hybrid contrastive learning module to minimize modality gap.

Specifically, given the ith visual emotion embedding $\mathbf{v}_i^r$ and jth acoustic emotion embedding $\mathbf{a}_j^r$, the popular cosine similarity is utilized to measure their semantic relevance

$$s\left(\mathbf{v}_i^r, \mathbf{a}_j^r\right) = \exp(\cos(\mathbf{v}_i^r, \mathbf{a}_j^r)). \tag{4}$$

1) Intramodal Contrastive Learning: It aims to learn the intramodal relationships between different emotional instances. That is, a positive pair is defined as the unimodal representations from different emotional samples of the same affective class, while a negative pair is defined as the unimodal representations from two emotional samples whose affective classes are different. Given the ith visual emotion embedding $\mathbf{v}_i^r$, we select a group of positive samples $\mathcal{P}_i^v$ and negative samples $\mathcal{N}_i^v$ according to their affective labels in a mini-batch of size B. Then, the intramodal contrastive loss of visual modality can be formulated as

$$\mathcal{L}_v^{\text{intra}} = -\sum_{i=1}^{B} \log \frac{\sum\limits_{\mathbf{v}_k^r \in \mathcal{P}_i^v} s\left(\mathbf{v}_i^r, \mathbf{v}_k^r\right)}{\sum\limits_{\mathbf{v}_k^r \in \mathcal{P}_i^v} s\left(\mathbf{v}_i^r, \mathbf{v}_k^r\right) + \sum\limits_{\mathbf{v}_j^r \in \mathcal{N}_j^v} s\left(\mathbf{v}_i^r, \mathbf{v}_j^r\right)}. \tag{5}$$

Often, the bidirectional learning scheme is able to improve the data utilization. Similarly, given the ith audio emotion embedding $\mathbf{a}_i$, we also select a group of positive samples $\mathcal{P}_i^a$ and negative samples $\mathcal{N}_i^a$ according to their affective labels in a mini-batch of size B. Accordingly, the intramodal contrastive loss of visual modality can be formulated as

$$\mathcal{L}_a^{\text{intra}} = -\sum_{i=1}^{B} \log \frac{\sum\limits_{\mathbf{a}_k^r \in \mathcal{P}_i^a} s\left(\mathbf{a}_i^r, \mathbf{a}_k^r\right)}{\sum\limits_{\mathbf{a}_k^r \in \mathcal{P}_i^a} s\left(\mathbf{a}_i^r, \mathbf{a}_k^r\right) + \sum\limits_{\mathbf{a}_j^r \in \mathcal{N}_j^v} s\left(\mathbf{a}_i^r, \mathbf{a}_j^r\right)}. \tag{6}$$

Accordingly, the intramodal contrastive loss of audio–visual modalities can be expressed as

$$\mathcal{L}_{\text{all}}^{\text{intra}} = \mathcal{L}_v^{\text{intra}} + \mathcal{L}_a^{\text{intra}}. \tag{7}$$

2) Cross-Modal Contrastive Learning: It aims to learn the cross-modal dynamic relationships between different emotional samples. That is, a positive pair is defined as the grouped multimodal emotional data from different modalities of the same affective class, while a negative pair is defined as the grouped multimodal emotional data from two heterogeneous samples whose affective classes are different. Given the visual emotion embedding $\mathbf{v}_i^r$ of an instance i, the positive samples from set $\mathcal{P}_i^a$ and negative samples from set $\mathcal{N}_i^a$ are selected for cross-modal contrastive learning. Then, the cross-modal contrastive loss of visual modality can be formulated as

$$\mathcal{L}_v^{\text{cross}} = -\sum_{i=1}^{B} \log \frac{\sum\limits_{\mathbf{a}_k^r \in \mathcal{P}_i^a} s\left(\mathbf{v}_i^r, \mathbf{a}_k^r\right)}{\sum\limits_{\mathbf{a}_k^r \in \mathcal{P}_i^a} s\left(\mathbf{v}_i^r, \mathbf{a}_k^r\right) + \sum\limits_{\mathbf{a}_j^r \in \mathcal{N}_j^a} s\left(\mathbf{v}_i^r, \mathbf{a}_j^r\right)}. \tag{8}$$

Similarly, given the audio emotion embedding $\mathbf{a}_i^r$ of an instance i, the positive samples from set $\mathcal{P}_i^v$ and negative samples from set $\mathcal{N}_i^v$ are selected. Then, the cross-modal contrastive loss of audio modality can be formulated as

$$\mathcal{L}_a^{\text{cross}} = -\sum_{i=1}^{B} \log \frac{\sum\limits_{\mathbf{v}_k^r \in \mathcal{P}_i^v} s\left(\mathbf{a}_i^r, \mathbf{v}_k^r\right)}{\sum\limits_{\mathbf{v}_k^r \in \mathcal{P}_i^v} s\left(\mathbf{a}_i^r, \mathbf{v}_k^r\right) + \sum\limits_{\mathbf{v}_j^r \in \mathcal{N}_j^v} s\left(\mathbf{a}_i^r, \mathbf{v}_j^r\right)}. \tag{9}$$

Accordingly, the cross-modal contrastive loss across audio–visual modalities can be expressed as

$$\mathcal{L}_{\text{all}}^{\text{cross}} = \mathcal{L}_v^{\text{cross}} + \mathcal{L}_a^{\text{cross}}. \tag{10}$$

3) Hard-Sample Contrastive Learning: The intramodal contrastive learning is performed to capture intra-modal instance relationships, while the cross-modal contrastive learning is designed to explore inter-class relationships. However, as the number of samples in these contrastive learning tasks increases, the risk of encountering invalid emotional data pairs also rises. To address this issue, we further parse a group of hard samples to guide the training process within the proposed framework, and thus promote the model to identify the abnormal emotion pairs more efficiently.

Specifically, we select the top-ranked k negative samples with the greatest similarity and the top-ranked k positive samples with the smallest similarity to regularize the learning process in a cross-modal learning way. Given a visual emotion embedding $\mathbf{v}_i^r$, we compute and sort the cross-modal similarity score of each pair in a mini-batch size, and then select k pairs that have low similarity from acoustic positive set $\mathcal{P}_i^a$ to form the hard positive set $\bar{\mathcal{P}}_i^a$

$$\bar{\mathcal{P}}_i^a = \underset{1,\ldots,k}{\mathtt{Rank}} \left\{ \mathtt{min} \left(s\left(\mathbf{v}_i^r, \mathbf{a}_w^r\right)_{w\in\mathcal{P}_i^a} \right) \right\}. \tag{11}$$

Similarly, we select k pairs that have high similarity from acoustic negative set $\mathcal{N}_i^a$ to form the hard negative set $\bar{\mathcal{N}}_i^a$

$$\bar{\mathcal{N}}_i^a = \underset{1,\ldots,k}{\mathtt{Rank}} \left\{ \mathtt{max} \left(s\left(\mathbf{v}_i^r, \mathbf{a}_w^r\right)_{w\in\mathcal{N}_i^a} \right) \right\}. \tag{12}$$

Since cross-emotion anomaly detection can be well regarded as a binary classification problem, their semantic-consistency labels can be generated naturally according to the semantic correspondence of emotional data pair. That is, if the affective expressions of audio–visual emotion data pair are matched, the values of these semantic-consistency labels are equal to 1, and 0 otherwise. Therefore, the hard-sample contrastive loss of visual modality can be derived as follows:

$$\mathcal{L}_v^{\text{hard}} = -\sum_{i=1}^{B} \left(\sum_{s\in\bar{\mathcal{P}}_i^a} \mathbf{y}_{i,s}^{v-a} \cdot \mathtt{log}(s(\mathbf{v}_i^r, \mathbf{a}_s^r)) + \sum_{m\in\bar{\mathcal{N}}_i^v} \mathbf{y}_{i,m}^{v-a} \cdot \mathtt{log}\left(s(\mathbf{v}_i^r, \mathbf{a}_m^r)\right) \right) \tag{13}$$

where $\mathbf{y}_{i,s}^{v-a}$ and $\mathbf{y}_{i,m}^{v-a}$ are respectively the affective-consistent labels for data pair $\{\mathbf{v}_i^r, \mathbf{a}_s^r\}$ and $\{\mathbf{v}_i^r, \mathbf{a}_m^r\}$, with value 1 for the semantic-consistency and 0 for the semantic-inconsistency.

Similarly, given an audio emotion embedding $\mathbf{a}_i^r$, we compute and sort the cross-modal similarity score of each pair in mini-batch, and then select k pairs that have low similarity from visual positive set $\mathcal{P}_i^v$ to form the hard positive set $\bar{\mathcal{P}}_i^v$

$$\bar{\mathcal{P}}_i^v = \underset{1,\ldots,k}{\mathtt{Rank}} \left\{ \mathtt{min} \left(s\left(\mathbf{a}_i^r, \mathbf{v}_w^r\right)_{w\in\mathcal{P}_i^v} \right) \right\}. \tag{14}$$

Similarly, we select k pairs that have high similarity from visual negative set $\mathcal{N}_i^v$ to form the hard negative set $\bar{\mathcal{N}}_i^v$

$$\bar{\mathcal{N}}_i^v = \underset{1,\ldots,k}{\mathtt{Rank}} \left\{ \mathtt{max} \left(s\left(\mathbf{a}_i^r, \mathbf{v}_w^r\right)_{w\in\mathcal{N}_i^v} \right) \right\}. \tag{15}$$

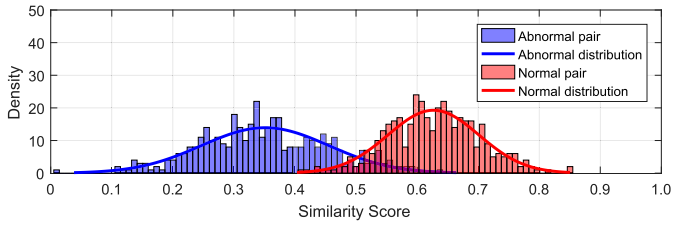

Fig. 3. Similarity distributions of semantic-consistency scores and semantic-inconsistency scores.

Therefore, the hard-sample contrastive loss of audio modality can be derived as follows:

$$\mathcal{L}_a^{\text{hard}} = -\sum_{i=1}^{B} \left(\sum_{s\in\bar{\mathcal{P}}_i^v} \mathbf{y}_{i,s}^{a-v} \cdot \mathtt{log}(s(\mathbf{a}_i^r, \mathbf{v}_s^r)) + \sum_{m\in\bar{\mathcal{N}}_i^v} \mathbf{y}_{i,m}^{a-v} \cdot \mathtt{log}\left(s(\mathbf{a}_i^r, \mathbf{v}_m^r)\right) \right) \tag{16}$$

where $\mathbf{y}_{i,s}^{a-v}$ and $\mathbf{y}_{i,m}^{a-v}$ are respectively the affective consistency label of emotional data pair $\{\mathbf{a}_i^r, \mathbf{v}_s^r\}$ and $\{\mathbf{a}_i^r, \mathbf{v}_m^r\}$, with value 1 for the semantic-consistency and 0 for the semantic-inconsistency. Accordingly, the total hard-sample contrastive loss can be obtained by

$$\mathcal{L}_{\text{all}}^{\text{hard}} = \mathcal{L}_v^{\text{hard}} + \mathcal{L}_a^{\text{hard}}. \tag{17}$$

The overall hybrid contrastive loss function in a mini-batch size is a weighted sum of shared embedding loss, intramodal contrastive loss, cross-modal contrastive loss, and hard-sample contrastive loss, which can be integrated as

$$\mathcal{L}_{\text{all}} = \mathcal{L}_{\text{share}} + \lambda_1 \mathcal{L}_{\text{all}}^{\text{intra}} + \lambda_2 \mathcal{L}_{\text{all}}^{\text{cross}} + \lambda_3 \mathcal{L}_{\text{all}}^{\text{hard}} \tag{18}$$

where λ_1, λ_2, and λ_3 are the hyper-parameters to balance the contributions of the different contrastive losses.

D. Semantic-Inconsistency Reasoning and Optimization

The main objective of SIR-HCL is to reason the semantic-inconsistency among the semantically irrelevant emotion pairs, while enhancing the semantic-consistency between semantically relevant ones. On the one hand, the emotions with semantic-inconsistency will be pushed away from each other, resulting in very small/large cross-modal similarity in the transformed feature space. On the other hand, the emotions with semantic-consistency will be pulled together, leading to very large cross-modal similarity in the transformed feature space. To illustrate this, we randomly group 300 normal emotion pairs and 300 abnormal emotion pairs from the MOSI dataset [26] to show their similarity differences. After the training process, the representative similarity distributions of semantic-consistency scores and semantic-inconsistency scores are shown in Fig. 3. It can be observed that the similarity scores of different emotional pairs are in different ranges. Inspired by this finding, we utilize two-component beta mixture model (BMM) [27] to fit the similarity distributions: $\mathbf{s}(i) = \mathtt{exp}(\mathtt{cos}(\mathbf{v}_i^r, \mathbf{a}_i^r))$ of each normal

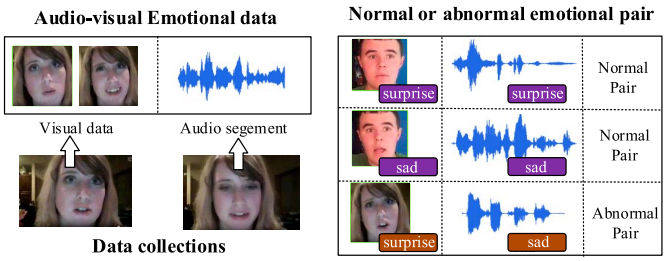

Fig. 4. Illustration of multimodal emotional data organizations.

TABLE I
ANNOTATIONS AND DESCRIPTIONS OF DATASETS

Datasets	Original	Redefine	Labels	c-Labels
MOSI	[−3,−2)	−3	Strong negative	Negative
	[−2,−1)	−2	Negative	
	[−1,0)	−1	Weak negative	
	0	0	Neutral	Neutral
MOSEI	(0,1]	1	Weak positive	Positive
	(1,2]	2	Positive	
	(2,3]	3	Strong positive	

and abnormal emotional pair, and thus obtain the probability of a sample pair being normal or abnormal as follows:

$$p_i = \sum_{m=1}^{M} \pi_m \texttt{Beta}\left(\mathbf{s}_i \mid \alpha_m, \beta_m\right) \tag{19}$$

where M is the total mixture number, π_m is the mth mixture coefficient, α_m and β_m are respectively the probability density function parameters of $\texttt{Beta}$ function for the mth mixture.

To initialize the model parameters, we fit the components of the BMM to the similarity scores of positive audio–visual emotion pairs and negative audio–visual emotion pairs during the training stage. The expectation-maximization (EM) algorithm is then employed to optimize the model parameters. During the testing phase, we compute the similarity score between visual–audio emotion pairs and utilize the derived probability distribution to determine whether the emotional state across audio-visual modalities is abnormal or not.

IV. Experiment

This section conducts a series of quantitative experiments on public multimodal sentiment datasets, and validates the effectiveness of the proposed SIR-HCL method. The experimental results, comparative analyses, and quantitative evaluations are detailed in the following sections.

A. Datasets and Implementations

In the experiments, two popular multimodal sentiment datasets, CMU-MOSI and CMU-MOSEI, are selected for evaluation, and their brief descriptions are clarified as follows.

1) *CMU-MOSI:* This dataset refers to the multimodal corpus of sentiment intensity [26], which consists of 2199 opinion video clips. Each opinion video is annotated with a sentiment score ranging from –3 to 3, which respectively corresponds to highly negative (–3), negative (–2), weakly negative (–1), neutral (0), weakly positive (+1), positive (+2), and highly positive (+3) to the sentiment intensity. The dataset is rigorously annotated with labels for subjectivity, sentiment intensity, per-frame and per-opinion annotated visual features, and per-millisecond audio features.

2) *CMU-MOSEI:* It is a large multimodal sentiment analysis and gender-balanced emotional dataset [28], consisting of 22 856 videos derived from 5000 videos and 1000 distinct speakers [28]. Each video inherently contains three modalities: visual, audio and text, and the visual and audio modalities are particularly employed in our experiments. Similarly, each sample is annotated by human annotators with a continuous sentiment score ranging from –3 to 3. In our work, we utilize 16 285 utterances for training, and 4643 utterances for testing.

These two datasets include diverse speakers with approximately equal gender distribution. The training, validation, and test sets are divided into a 3:1:1 ratio, where the emotional segments are randomly selected from the same video. Similar to work [29], we extract 35-dimensional visual features, primarily comprising basic and advanced facial action units, from video frames using FACET [30]. Additionally, 74-dimensional audio features, including 12 Mel-frequency cepstral coefficients (MFCCs) and other low-level acoustic features, are extracted from acoustic signals using COVAREP [31]. As shown in Fig. 4, the paired emotional instances derived from the same video clips share the same affective label. Since no abnormal multimodal emotional instances exist in the datasets, we randomly shuffle 50% of the audio–visual emotional data pairs from the same speaker to generate abnormal emotional pairs that are affectively mismatched. Following the intuitive sentiment categorization [26], as shown in Table I, the affective labels of the two datasets are ultimately processed into three categories: positive, neutral, and negative.

For these multimodal emotional datasets, the visual subnetwork and audio subnetwork are employed to extract visual emotion embeddings and acoustic emotion embeddings, respectively, each with a dimensionality of 256. Within these networks, the number of hidden neurons in each cell is set to 128, the dropout rate is fixed at 0.5, and the batch size is set to 16 for MOSI and 32 for MOSEI. The number k of hard samples is set to 6, and the parameters $\{\lambda_1, \lambda_2, \lambda_3\}$ are set to $\{1, 1, 0.5\}$. The proposed model is optimized using Stochastic gradient descent (SGD) with a learning rate of 0.0001. For the two-component beta mixture model estimation, the probability threshold is fixed at 0.5 to reason about abnormal emotions across different modalities. In all experiments, the learning rate is decayed by 0.1 after 10 epochs. The entire network is trained in an end-to-end manner, and the network parameters are updated through backpropagation.

B. Evaluation Metrics and Baseline Methods

The goal of cross-emotion anomaly detection is to identify abnormal multimodal emotion pair whose semantic

patterns are inconsistent across different modalities. This task can be formulated as a typical binary identification problem. Therefore, the popular true positive rate (TPR) and false positive rate (FPR) are selected for quantitative analysis: $\mathtt{TPR} = (\mathtt{TP}/\mathtt{TP} + \mathtt{FN})$, $\mathtt{FPR} = (\mathtt{FP}/\mathtt{TN} + \mathtt{FP})$, where TP, FN, TN, and FP, respectively, represent the number of true positives, false negatives, true negatives, and false positives. For the performance evaluation, the larger TPR values often reveal the better identification performance, while the smaller FPR values indicate the better detection results [32]. Additionally, accuracy $\mathtt{Accuracy} = (\mathtt{TP} + \mathtt{TN}/\mathtt{TP} + \mathtt{TN} + \mathtt{FP} + \mathtt{FN})$ and AUC values are also selected to validate the detection performances.

The proposed framework is the first attempt to detect anomalous emotion pairs across audio-visual emotional data, and there are no relevant works to tackle this problem. For meaningful comparisons, we extend a few relevant methods to perform cross-emotion anomaly detection tasks. MISA [29] learns modality-invariant and specific representations for multi-modal sentiment analysis. CMAD [20] exploits a deep structured framework to characterize the feature representations between heterogeneous data samples, and utilizes a threshold to detect the abnormal examples across different modalities. NCR [33] employs triplet loss and Gaussian mixture distributions to distinguish different kinds of data pairs. CMPC [7] utilizes cross-modal prototype contrastive learning to perform voice-face matching in a cross-modal way. HFIR [17] employs information relevance as the matching degree between cross-modal features at the emotional semantic level, and utilizes hybrid fusion based on information relevance (HFIR) for multimodal sentiment analysis. Since CMPC and MISA do not directly detect anomalous pairs, we also utilize a two-component Beta Mixture Model as stated within the proposed framework to distinguish possible anomalous emotion samples. For the other baselines, we utilize the same similarity threshold value to detect the possible abnormal emotions across different modalities.

C. Performance Comparison and Analysis

1) Results of Detection Performance: The cross-emotion anomaly detection results obtained by different methods and tested on different datasets are shown in Tables II and III, respectively. It can be seen that the proposed method has delivered very competitive cross-emotion anomaly detection performances, and outperforms most of the baselines in different datasets. For the smaller CMU-MOSI dataset, the abnormal emotion detection results obtained by the proposed SIR-HCL approach do not differ significantly from the baseline methods. The main reason lies that the CMU-MOSI dataset has fewer examples, and the emotional complexity is not very competitive. Accordingly, the detection results obtained by different methods do not differ very much. Note that, the proposed SIR-HCL method has delivered better AUC, FPR, and TPR scores. For instance, the TPR value obtained by the proposed approach reached up to 0.9563. This indicates that the proposed SIR-HCL approach holds a strong ability to detect the abnormal emotions across different modalities.

TABLE II
AUC, FPR, TPR, AND ACCURACY RESULTS EVALUATED ON MOSI DATASET

Method	AUC	FPR	TPR	Accuracy
CMAD [20]	0.8437	0.2230	0.9193	0.8437
CMPC [7]	0.8627	0.2471	0.7954	0.7621
NCR [33]	0.8342	0.2189	0.8370	0.8219
MISA [29]	0.8549	0.1653	0.8739	0.8739
HFIR [17]	0.9079	0.1823	0.8635	0.8817
SIR-HCL	**0.9143**	**0.1531**	**0.9468**	**0.8876**

Note: The best results are highlighted in bold.

TABLE III
AUC, FPR, TPR, AND ACCURACY RESULTS EVALUATED ON MOSEI DATASET

Method	AUC	FPR	TPR	Accuracy
CMAD [20]	0.8734	0.1781	0.8507	0.8767
CMPC [7]	0.8865	0.1939	0.9312	0.8419
NCR [33]	0.8627	0.1782	0.9234	0.8291
MISA [29]	0.8754	0.1887	0.9155	0.8471
HFIR [17]	0.8971	0.2135	0.9228	0.8651
SIR-HCL	**0.9137**	**0.1723**	**0.9563**	**0.8895**

Note: The best results are highlighted in bold.

For the large CMU-MOSI dataset, it can be found that the competing baselines have delivered relatively lower AUC, TPR, and accuracy values, while generating larger FPR values. For instance, the AUC score and accuracy values obtained by the MISA method are respectively equal to 0.8754 and 0.8471, while the AUC score and accuracy value obtained by the HFIR method are respectively equal to 0.8971 and 0.8651. Notably, these two methods utilize the fusion model to bridge the semantic gap between heterogeneous emotion features, which can detect some obvious abnormal emotion pairs. However, these methods often fail to detect inconsistent emotional behaviors when the abnormal emotions exhibit minor differences. Specifically, NCR [33] just considers one positive sample and one negative sample of the specified instance, which therefore cannot learn the discriminative latent embeddings and therefore result in a lower performance. CMPC [7] employs the unsupervised clustering to construct the positive embedding and negative embedding between voice-face representations, and the accuracy score obtained by this approach is 0.8419. Note that, this approach ignores the intra-modal negative samples to explore the fine-grained representations in each modality, and its performance is uncompetitive when processing the large-scale emotional dataset. CMAD [20] first utilizes a deep structured framework to learn the feature representations between heterogeneous modalities, and then applies a simple anomaly threshold to identify the anomalies whose patterns are significantly disparate across different modalities. Remarkably, this

TABLE IV
CROSS-EMOTION ANOMALY DETECTION RESULTS OBTAINED BY DIFFERENT APPROACHES AND TESTED ON DIFFERENT DATASETS

Datasets	Method	Undevided		Positive-Negative		Positive-Neutral		Neutral-Negative	
		Acc	AUC	Acc	AUC	Acc	AUC	Acc	AUC
CMU-MOSI	CMAD [20]	0.8437	0.8859	0.8679	0.854	0.8211	0.8251	0.8367	0.825
	CMPC [7]	0.7621	0.8627	0.8123	0.850	0.8216	0.8438	0.8246	0.8317
	NCR [33]	0.8219	0.8342	0.8874	0.8421	0.8168	0.8514	0.8722	0.8861
	MISA [29]	0.8739	0.8546	0.8464	0.8320	0.8129	0.8764	0.8905	0.8910
	HFIR [17]	0.8817	0.9079	0.8852	0.8751	**0.8974**	0.9094	0.9076	0.8861
	SIR-HCL	**0.8876**	**0.9143**	**0.8982**	**0.9135**	0.8937	**0.9178**	**0.9107**	**0.9165**
CMU-MOSEI	CMAD [20]	0.8767	0.8734	0.8511	0.8692	0.8397	0.8463	0.8712	0.8609
	CMPC [7]	0.8419	0.8865	0.8733	0.8512	0.8250	0.8137	0.7882	0.8215
	NCR [33]	0.8291	0.8627	0.8456	0.8419	0.8845	0.8512	0.8142	0.8736
	MISA [29]	0.8471	0.8754	0.8367	0.8329	0.8190	0.8268	0.8506	0.8578
	HFIR [17]	0.8651	0.8971	0.8978	0.8879	0.8651	0.8539	0.8613	0.8426
	SIR-HCL	**0.8895**	**0.9137**	**0.9106**	**0.8958**	**0.8976**	**0.8902**	**0.8913**	**0.8998**

Note: The best results are highlighted in bold.

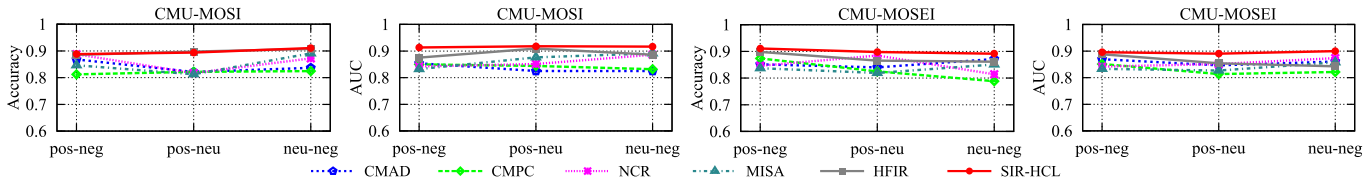

Fig. 5. Detection performance obtained by different approaches on different datasets. The abbreviation and corresponding full form of each label are as follows: pos, positive; neu, neutral; neg, negative.

method just utilizes the triple loss to penalize the instances with inconsistent pairs, failing to leverage the negative samples that are crucial for identifying abnormal emotions. As a result, its detection performance needs further improvement. Comparatively speaking, the proposed SIR-HCL approach can well measure the semantic-consistency among normal emotion pairs, while accurately identifying semantic-inconsistency in abnormal emotion pairs. Consequently, SIR-HCL consistently achieves higher cross-emotion anomaly detection accuracy than the competing baselines.

Further, we detail the cross-emotion anomaly detection tasks into three subtasks: positive–negative anomaly, positive–neutral anomaly, and negative–neutral anomaly. As shown in Table IV, it can be observed that the proposed SIR-HCL approach consistently achieves better detection performance under different abnormal conditions. For the more challenging negative–neutral task, the detection performance obtained by the baselines methods often yields relatively lower scores. For the CMU-MOSEI dataset, the proposed SIR-HCL method yields about 2.01% improvement on recognizing "neutral–negative" abnormal emotion pairs compared with CMAD method [20]. Fig. 5 shows the comparison curves obtained by different approaches. It can be found that the accuracy values evaluated on "positive–negative" abnormal emotion pairs and obtained by competing baselines were relatively unstable when tested on MOSI dataset, In contrast to this, our proposed SIR-HCL method demonstrates very stable performance on both datasets, and the corresponding accuracy values are always higher than the results obtained by all the competing baselines. That is, the proposed SIR-HCL approach not only effectively handles various abnormal emotion detection tasks across different modalities, but also delivers relatively stable detection performance under different abnormal conditions.

Besides, we evaluate the proposed methods on fine-grained cross-emotion anomaly detection tasks, where the affective labels are defined in a more granular manner, i.e., strong negative (s-neg), negative, weak negative (w-neg), neutral, weak positive (w-pos), positive and strong positive (s-pos). Accordingly, five abnormal cross-emotion cases are designed for enhanced fine-grained detection tasks, i.e., negative–positive, negative–neural, positive–neural, weak-positive to weak-negative and strong-positive to strong-negative. As shown in Table V, it can be observed that CMAD and NCR methods achieve relatively lower ACC and AUC scores, while CMPC, MISA, and HFIR approaches also exhibit degraded performance across weak-positive and weak-negative data collections. By contrast, as shown in Fig. 6, our proposed SIR-HCL method demonstrates very competitive detection performances and significantly outperforms these baseline methods. This indicates that the proposed framework is capable of detecting abnormal emotional pairs even when the affective differences are subtle and the cross-emotion scenarios are highly complex.

2) Ablation Study: Within the proposed framework, the designed hybrid contrastive loss plays a critical role in learning

TABLE V
ENHANCED FINE-GRAINED CROSS-EMOTION ANOMALY DETECTION PERFORMANCE ON CMU-MOSEI DATASET

Methods	Pos-Neg		Pos-Neu		Neu-Neg		Neg(w)-Pos(w)		Neg(s)-Pos(s)	
	Acc	AUC	Acc	AUC	Acc	AUC	Acc	AUC	Acc	AUC
CMAD [20]	0.7951	0.7178	0.7820	0.7865	0.8448	0.8589	0.7111	0.7112	0.7069	0.7077
CMPC [7]	0.8352	0.8380	0.8465	0.8513	0.8396	0.850	0.7922	0.7801	0.8426	0.8397
NCR [33]	0.7065	0.7246	0.7114	0.7295	0.8569	0.8347	0.7514	0.7517	0.8269	0.8310
MISA [29]	0.8093	0.8119	0.8307	0.8482	0.8298	0.8313	0.7815	0.7780	0.8223	0.8377
HFIR [17]	0.8436	0.8529	0.8627	0.8775	0.8392	0.8491	0.8507	**0.8374**	0.8782	0.8834
SIR-HCL	**0.8471**	**0.8981**	**0.8753**	**0.8907**	**0.8816**	**0.8891**	**0.8652**	0.8317	**0.8857**	**0.8891**

Note: The best results are highlighted in bold.

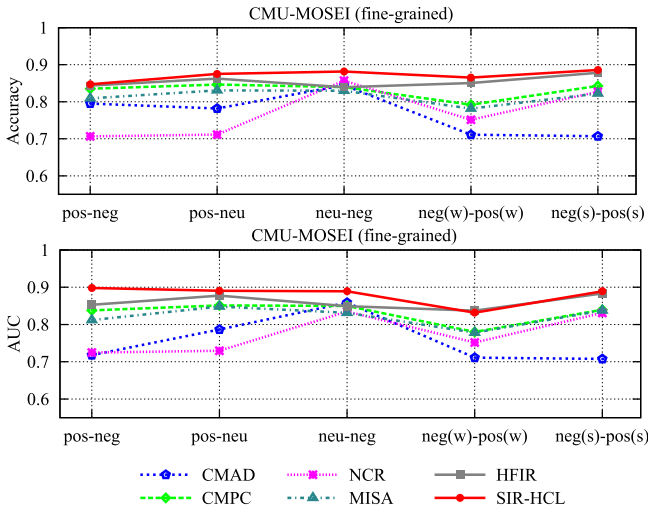

Fig. 6. Enhanced fine-grained detection performance obtained by different approaches on CMU-MOSEI dataset.

the discriminative cross-emotion embeddings. As illustrated in Table VI, we conduct an ablation study to evaluate the contribution of each loss component by sequentially removing it from the model, in which w/o-$\mathcal{L}_{\text{share}}$, w/o-$\mathcal{L}_{\text{intra}}$, w/o-$\mathcal{L}_{\text{cross}}$, and w/o-$\mathcal{L}_{\text{hard}}$, respectively, denote the removal of shared embedding loss, intra-modal contrastive loss, cross-modal contrastive loss and hard-sample contrastive loss. It can be found that the removal of shared embedding loss results in a slight performance degradation, which indicates that the shared residual block can well bridge the modality gap between heterogeneous emotions. Meanwhile, the learning of cross-emotion relationships is a fundamental component that leads to the high abnormal emotion detection performance, while the intra-modal contrastive loss and hard-sample contrastive loss also contribute to boosting the detection performance. This demonstrates that the proposed model effectively enhances the discriminative power of audio–visual embeddings, achieving significantly better results when the intra-modal, cross-modal, and hard-sample contrastive losses are jointly incorporated. Overall, the integration of shared embedding loss, intramodal contrastive loss, cross-modal contrastive loss, and hard-sample contrastive loss enables the learning of more discriminative cross-emotion

TABLE VI
ABLATION STUDIES ON THE CMU-MOSEI DATASET

Method	MOSE		MOSEI	
	AUC	Accuracy	AUC	Accuracy
w/o-$\mathcal{L}_{\text{share}}$	0.8573	0.8669	0.8792	0.8753
w/o-$\mathcal{L}_{\text{intra}}$	0.8705	0.8613	0.8727	0.8681
w/o-$\mathcal{L}_{\text{cross}}$	0.8692	0.8718	0.8691	0.8781
w/o-$\mathcal{L}_{\text{hard}}$	0.8737	0.8715	0.8893	0.8703
$\mathcal{L}_{\text{all}}$	**0.9143**	**0.8876**	**0.9137**	**0.8875**

Note: The best results are highlighted in bold.

embeddings, thereby significantly improving cross-emotion anomaly detection performance.

3) Visualization and Analysis: To visually verify the superiority of the proposed SIR-HCL model, Fig. 7 shows several representative cross-emotion anomaly detection examples, demonstrating the effectiveness of the proposed SIR-HCL framework. To be specific, the detection results marked in green indicate that the facial expression and voice clip share the same affective class, and the detection results are recognized as the normal pair. In contrast to this, the detection results marked in red indicate that the facial expression and voice clip do not belong to the same affective class, and the detection results are recognized as the abnormal pair. It can be observed that the derived similarity scores provide an intuitive measure of the correlation degree between heterogeneous emotional modalities. That is, if the facial expression and voice clip share the same affective class, the proposed model is able to recognize their strong relevance. Conversely, if the facial expression and voice clip are mistakenly grouped together, the proposed SIR-HCL framework is capable of identifying such irrelevance and its corresponding similarity score is very small. This indicates that the proposed model exhibits high discriminability to reason the affective relationship between heterogeneous emotional samples, making it highly effective in detecting abnormal emotions practically.

Further, we utilize the t-SNE algorithm to visualize the derived multimodal emotional embedding vectors. As shown in Fig. 8, it can be found that the initial embeddings of audio–visual emotions belonging to the same affective class are separated into two distinct clusters due to modality gap.

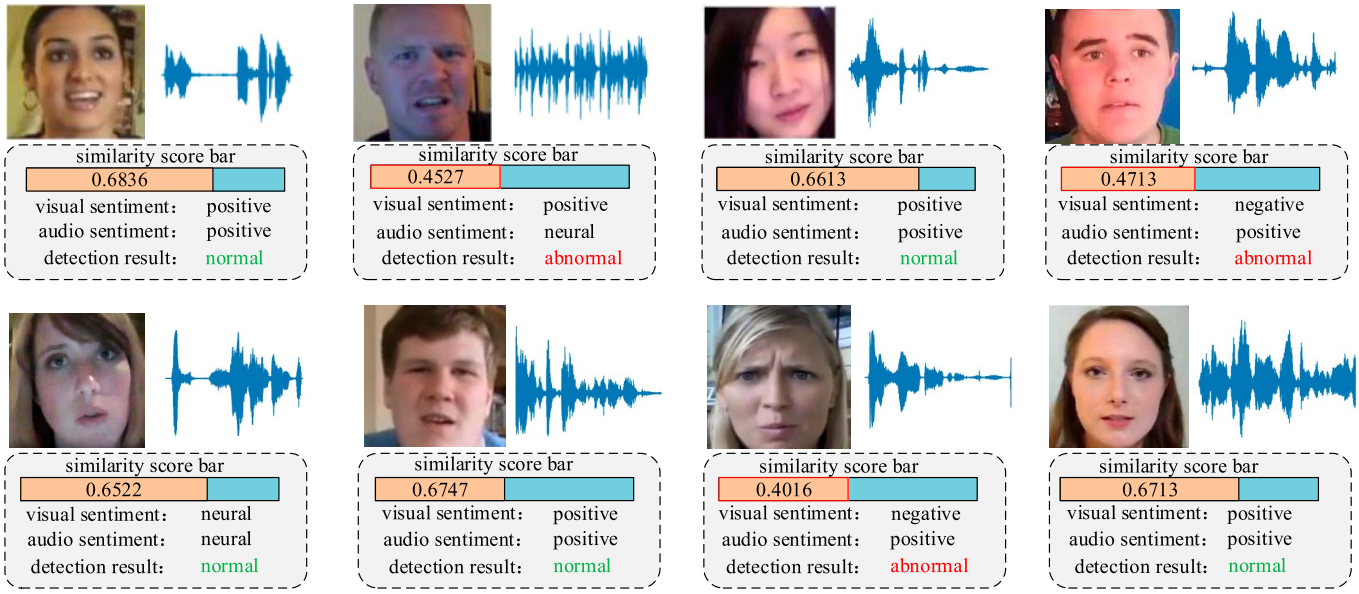

Fig. 7. Visualization of representative cross-emotion anomaly detection examples. For each audio-visual emotion pair, a facial expression frame and an audio clip respectively represent a kind of affective class. The detection results marked in red (abnormal pair) indicate that the facial expression and voice clip do not behave consistently with each other, while the detection results marked in green (normal pair) indicate that the facial expression and voice clip behave consistently with each other.

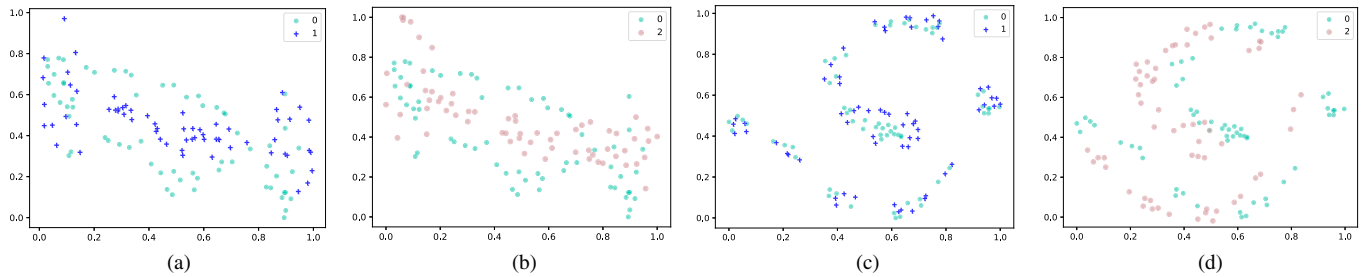

Fig. 8. t-SNE visualization of the cross-modal embeddings in the case without or with training. The green color represents visual modality, while the blue and pink colors respectively represent audio modality with the same or different affective classes. (a) Positive pairs before training. (b) Negative pairs before training. (c) Positive pairs after training. (d) Negative pairs after training.

Moreover, the emotional distributions from the different affective classes are always overlapping. Under such circumstances, it is very difficult to reason the anomalous emotion pairs with semantic-inconsistency. After training, the audio–visual emotions belonging to the same affective class are clustered closely together, while those from different affective classes are clearly separated. The main reason lies that the proposed SIR-HCL model is able to aggregate the cross-emotion data pairs of the same affective class close together, while pulling those of different affective classes away. As a result, the derived cross-emotion embeddings are semantically meaningful, to enhance abnormal emotion detection performances.

Besides, we investigate the similarity distributions of abnormal and normal emotional data pairs predicted by the proposed SIR-HCL model at different learning stages. Fig. 9 shows the similarity distributions of 600 abnormal and 600 normal emotion pairs derived from the MOSEI dataset. On the one hand, it can be observed that the overlapping regions of similarity distributions before training are significantly larger than those after training. On the other hand, the similarity distributions of abnormal emotion pairs exhibit a wide range before training, with some peaks of their distribution waves often falling within similar intervals. Under such circumstances, it is very difficult to identify the anomalous emotional data pairs across audio-visual emotion data. After training, it can be seen that most similarity scores for normal emotional pairs are significantly higher than those for abnormal pairs, and the main peaks of their distribution waves are clearly separated into distinct intervals. This clear distinction between the similarity distributions can be effectively utilized to differentiate abnormal emotional data pairs. This demonstrates that the proposed model provides valuable cross-modal information to reason about semantic inconsistency, thereby enabling the detection of anomalous emotional samples across audio-visual data.

V. Conclusion

Cross-emotion anomaly detection across heterogeneous modalities is a relatively emerging topic in the field of

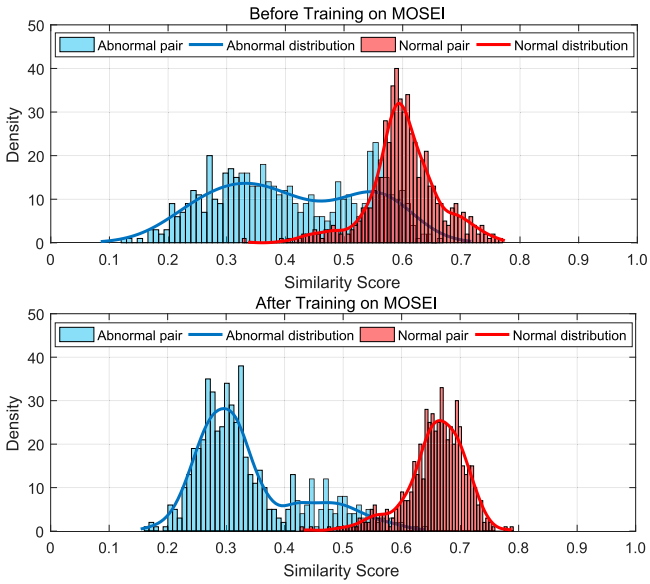

Fig. 9. Similarity distributions of abnormal and normal emotional data pairs derived before and after training processes.

multimodal sentiment analysis. This work presents an efficient cross-emotion anomaly detection approach via semantic-inconsistency reasoning and hybrid contrastive learning, which effectively identifies anomalous emotion pairs across audio–visual data. Within the proposed framework, an innovative hybrid contrastive learning approach is designed to enlarge the semantic inconsistency between abnormal emotional data pairs from different affective classes, while strengthening the semantic correspondence and feature correlation between normal emotional data pairs from the same affective class. Additionally, a bidirectional learning scheme is employed to enhance data utilization, and a two-component BMM is utilized to reason about anomalous emotion pairs with semantic inconsistency in a more interpretable manner. Extensive experiments have shown its competitive performance over the state of the arts.

REFERENCES

[1] X. Jin, P. Jing, J. Wu, J. Xu, and Y. Su, "Visual sentiment classification via low-rank regularization and label relaxation," *IEEE Trans. Cogn. Develop. Syst.*, vol. 14, no. 4, pp. 1678–1690, Dec. 2022.

[2] C. Park et al. "The neural systems of emotion regulation and abnormalities in major depressive disorder," *Behav. Brain Res.*, vol. 367, pp. 181–188, 2019.

[3] B. Yang, J. Cao, N. Wang, and X. Liu, "Anomalous behaviors detection in moving crowds based on a weighted convolutional autoencoder-long short-term memory network," *IEEE Trans. Cogn. Develop. Syst.*, vol. 11, no. 4, pp. 473–482, Dec. 2019.

[4] W. Liu, J.-L. Qiu, W.-L. Zheng, and B.-L. Lu, "Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition," *IEEE Trans. Cogn. Develop. Syst.*, vol. 14, no. 2, pp. 715–729, Jun. 2022.

[5] X. Sun, C. Zhang, and L. Li, "Dynamic emotion modelling and anomaly detection in conversation based on emotional transition tensor," *Inf. Fusion*, vol. 46, pp. 11–22, 2019.

[6] T. Sorensen, E. Zane, T. Feng, S. Narayanan, and R. Grossman, "Cross-modal coordination of face-directed gaze and emotional speech production in school-aged children and adolescents with ASD," *Sci. Rep.*, vol. 9, no. 1, 2019, Art. no. 18301.

[7] B. Zhu et al., "Unsupervised voice-face representation learning by cross-modal prototype contrast," in *Proc. Int. Joint Conf. Artif. Intell.*, 2022, pp. 3787–3794.

[8] A. Jan, H. Meng, Y. F. B. A. Gaus, and F. Zhang, "Artificial intelligent system for automatic depression level analysis through visual and vocal expressions," *IEEE Trans. Cogn. Develop. Syst.*, vol. 10, no. 3, pp. 668–680, Sep. 2018.

[9] W. Yu, H. Xu, Z. Yuan, and J. Wu, "Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 10790–10797.

[10] L. T. Germine, L. Garrido, L. Bruce, and C. Hooker, "Social anhedonia is associated with neural abnormalities during face emotion processing," *Neuroimage*, vol. 58, no. 3, pp. 935–945, 2011.

[11] C. Clavel, L. Devillers, G. Richard, I. Vasilescu, and T. Ehrette, "Detection and analysis of abnormal situations through fear-type acoustic manifestations," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2007, pp. 21–24.

[12] S. Gannouni, A. Aledaily, K. Belwafi, and H. Aboalsamh, "Emotion detection using electroencephalography signals and a zero-time windowing-based epoch estimation and relevant electrode identification," *Sci. Rep.*, vol. 11, no. 1, 2021, Art. no. 7071.

[13] J. Zhu, F. Deng, J. Zhao, D. Liu, and J. Chen, "UAED: Unsupervised abnormal emotion detection network based on wearable mobile device," *IEEE Trans. Netw. Sci. Eng.*, vol. 10, no. 6, pp. 3682–3696, Nov./Dec. 2023.

[14] G. Sharma, A. Dhall, and J. Cai, "Audio-visual automatic group affect analysis," *IEEE Trans. Affect. Comput.*, vol. 14, no. 2, pp. 1056–1069, Apr./Jun. 2023.

[15] F. Alvarez, M. Popa, V. Solachidis, and G. Hernández-Peñaloza, "Behavior analysis through multimodal sensing for care of parkinson's and alzheimer's patients," *IEEE MultiMedia*, vol. 25, no. 1, pp. 14–25, Jan./Mar. 2018.

[16] T. Horii, Y. Nagai, and M. Asada, "Modeling development of multimodal emotion perception guided by tactile dominance and perceptual improvement," *IEEE Trans. Cogn. Develop. Syst.*, vol. 10, no. 3, pp. 762–775, Sep. 2018.

[17] D. Chen, W. Su, P. Wu, and B. Hua, "Joint multimodal sentiment analysis based on information relevance," *Inf. Process. Manage.*, vol. 60, no. 2, 2023, Art. no. 103193.

[18] A. Dawel, R. O'Kearney, E. McKone, and R. Palermo, "Not just fear and sadness: Meta-analytic evidence of pervasive emotion recognition deficits for facial and vocal expressions in psychopathy," *Neurosci. Biobehav. Rev.*, vol. 36, no. 10, pp. 2288–2304, 2012.

[19] Y. Jiang, K. Hirota, Y. Dai, Y. Ji, and S. Shao, "Abnormal emotion recognition based on audio-visual modality fusion," in *Proc. Int. Conf. Intell. Robot. Appl.*, 2023, pp. 162–173.

[20] Y. Li, N. Liu, J. Li, M. Du, and X. Hu, "Deep structured cross-modal anomaly detection," in *Proc. Int. Joint Conf. Neural Netw.*, 2019, pp. 1–8.

[21] X. Liu, Y. He, Y.-M. Cheung, X. Xu, and N. Wang, "Learning relationship-enhanced semantic graph for fine-grained image-text matching," *IEEE Trans. Cybern.*, vol. 54, no. 2, pp. 948–961, Feb. 2024.

[22] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, "A survey on contrastive self-supervised learning," *Technologies*, vol. 9, no. 1, p. 2, 2021.

[23] M. Kim, J. Tack, and S. J. Hwang, "Adversarial self-supervised contrastive learning," in *Proc. 34th Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 2983–2994.

[24] S. Yang, B. Linares-Barranco, Y. Wu, and B. Chen, "Self-supervised high-order information bottleneck learning of spiking neural network for robust event-based optical flow estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, 2024, doi: 10.1109/TPAMI.2024.3510627.

[25] L. Wu, H. Lin, C. Tan, Z. Gao, and S. Z. Li, "Self-supervised learning on graphs: Contrastive, generative, or predictive," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 4, pp. 4216–4235, Apr. 2023.

[26] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages," *IEEE Intell. Syst.*, vol. 31, no. 6, pp. 82–88, Nov./Dec. 2016.

[27] Z. Ma and A. Leijon, "Bayesian estimation of beta mixture models with variational inference," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 11, pp. 2160–2173, Nov. 2011.

[28] A. Bagher Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2236–2246.

[29] D. Hazarika, R. Zimmermann, and S. Poria, "MISA: Modality-invariant and -specific representations for multimodal sentiment analysis," in *Proc. ACM Int. Conf. Multimedia*, 2020, pp. 1122–1131.
[30] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: An open source facial behavior analysis toolkit," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2016, pp. 1–10.
[31] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "COVEREP- A collaborative voice analysis repository for speech technologies," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 960–964.
[32] X. Liu, Z. Hu, H. Ling, and Y.-M. Cheung, "MTFH: A matrix tri-factorization hashing framework for efficient cross-modal retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 964–981, Mar. 2021.
[33] Z. Huang et al., "Learning with noisy correspondence for cross-modal matching," in *Proc. Int. Adv. Neural Inf. Process. Syst.*, 2021, pp. 29406–29419.

Xin Liu (Senior Member, IEEE) received the Ph.D. degree in computer science from Hong Kong Baptist University, Hong Kong, SAR, China, in 2013.

From 2017 to 2018, he was a Visiting Scholar with Computer & Information Sciences Department, Temple University, Philadelphia, PA, USA. Currently, he is a Full Professor with the Department of Computer Science and Technology, Huaqiao University, Xiamen, China, and also a Research Fellow with the Department of Computer Science, Hong Kong Baptist University. His research interests include multimedia data analysis, pattern recognition, and affective computing.

Qiyan Chen received the B.S. degree in computer science and technology from Xiamen University of Technology, Xiamen, China, in 2022.

Currently, she is a Research Fellow with Xiamen Key Laboratory of Computer Vision and Pattern Recognition, Xiamen, China, and also a Research Fellow with Fujian Key Laboratory of Big Data Intelligence and Security, Xiamen. Her research interests include multimedia content analysis, pattern recognition, and deep learning.

Yiu-ming Cheung (Fellow, IEEE) received the Ph.D. degree in computer science from the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, China, in 2000.

Currently, he is a Chair Professor (Artificial Intelligence) with the Department of Computer Science, Hong Kong Baptist University, Hong Kong, SAR, China. His research interests include machine learning, pattern recognition, and visual computing.

Dr. Cheung is the Editor-in-Chief (2023) of IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTATIONAL INTELLIGENCE, and serves as an Associate Editor for IEEE TRANSACTIONS ON COGNITIVE AND DEVELOPMENTAL SYSTEMS, IEEE TRANSACTIONS ON CYBERNETICS, *Pattern Recognition, Knowledge and Information Systems*, to name a few. He is an IET Fellow, AAAS Fellow, and BCS Fellow. For details, see: https://www.comp.hkbu.edu.hk/ymc.

Shu-Juan Peng (Senior Member, IEEE) received the Ph.D. degree in computer science from Wuhan University, Wuhan, China, in 2009.

Currently, she is a Full Professor with the Department of Artificial Intelligence, Huaqiao University, Xiamen, China, and also a Research Fellow with the Key Laboratory of Computer Vision and Machine Learning (Huaqiao University), Fujian Province University, Xiamen. Her research interests include multimedia data analysis and pattern recognition.