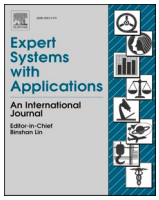
Expert Systems With Applications 237 (2024) 121692

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



Review

Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects

Shiqing Zhang [a,*], Yijiao Yang [a], Chen Chen [a], Xingnan Zhang [a], Qingming Leng [b], Xiaoming Zhao [a]

[a] Institute of Intelligent Information Processing, Taizhou University, Taizhou 318000, Zhejiang, China
[b] School of Electronic and Information Engineering, Jiujiang University, Jiujiang 332005, China

ARTICLE INFO

Keywords:
Multimodal emotion recognition
Deep learning
Feature extraction
Multimodal information fusion, review

ABSTRACT

Emotion recognition has recently attracted extensive interest due to its significant applications to human–computer interaction. The expression of human emotion depends on various verbal and non-verbal languages like audio, visual, text, *etc*. Emotion recognition is thus well suited as a multimodal rather than single-modal learning problem. Owing to the powerful feature learning capability, extensive deep learning methods have been recently leveraged to capture high-level emotional feature representations for multimodal emotion recognition (MER). Therefore, this paper makes the first effort in comprehensively summarize recent advances in deep learning-based multimodal emotion recognition (DL-MER) involved in audio, visual, and text modalities. We focus on: (1) MER milestones are given to summarize the development tendency of MER, and conventional multimodal emotional datasets are provided; (2) The core principles of typical deep learning models and its recent advancements are overviewed; (3) A systematic survey and taxonomy is provided to cover the state-of-the-art methods related to two key steps in a MER system, including feature extraction and multimodal information fusion; (4) The research challenges and open issues in this field are discussed, and promising future directions are given.

1. Introduction

Emotional interaction is a basic physical and psychological need in human's daily life. People express their emotions in different ways, which always correspond to different behaviors (Narayanan & Georgiou, 2013; Vinciarelli et al., 2011). With the widespread popularity of social media, online shopping and education, interactive video games, *etc*., the demand of accurately identifying human emotions, namely, emotion recognition, is greatly increased in recent years (Wang & Zhao, 2022), due to its significant applications to human–computer interaction (HCI) (Chowdary, Nguyen, & Hemanth, 2021), emotion-aware recommendation (Qian, Zhang, Ma, Yu, & Peng, 2019), e-learning environments (Bahreini, Nadolski, & Westera, 2016), immersive virtual reality (Marín-Morales, Llinares, Guixeres, & Alcañiz, 2020), *etc*.

Emotion recognition is a definitely challenging issue owing to the complexity of human emotion expression. People frequently use various verbal and non-verbal languages, including speech/audio signal, facial expression, text and so on, so as to express their emotions. A significant corpus of single-modal works has developed emotion recognition research, such as speech/audio, visual, and text emotion recognition. Although these single-modal works usually obtain promising performance, different modalities can be highly correlated with the same emotion expression. In this case, integrating multiple modalities is very beneficial to promote the performance of emotion recognition. In addition, according to the advances in understanding the nature of emotion expression in a basic emotion theory (Keltner, Sauter, Tracy, & Cowen, 2019), emotion expression is multimodal dynamic patterns of behaviors involved in not only facial muscle movements but also variations in gaze, body movements, gestures, hand movements, the voice, and so on. In essence, emotion recognition is thus more suitable to be regarded as a problem of multimodal emotion recognition (MER) (Bänziger, Grandjean, & Scherer, 2009).

* Corresponding author.
E-mail addresses: tzczsq@163.com (S. Zhang), yang_yijiao@163.com (Y. Yang), beans552@163.com (C. Chen), zhangxingnan11@163.com (X. Zhang), qingming_leng@jju.edu.cn (Q. Leng), tzxyzxm@163.com (X. Zhao).

https://doi.org/10.1016/j.eswa.2023.121692
Received 14 February 2023; Received in revised form 2 September 2023; Accepted 16 September 2023
Available online 18 September 2023

With the recent advancement of deep learning algorithms (LeCun, Bengio, & Hinton, 2015; Schmidhuber, 2015), deep learning based emotion recognition techniques has exhibited inspiring performance. In particular, with the aid of a multi-layer network structure, deep learning techniques are capable of automatically learning high-level feature representations for emotion recognition, thereby frequently outperforming hand-crafted features based emotion recognition methods (Ozseven, 2023). The above-mentioned two aspects motivate us to conduct a comprehensive survey for deep learning based multimodal emotion recognition (DL-MER) and discuss several future directions.

Though some surveys have also summarized the developments of MER or deep learning for emotion recognition in different aspects, as shown in Table 1. Only three surveys by Latha *et al.*, 2016 (Latha & Priya, 2016), Gu *et al.*, (Gu et al., 2021) and Abdullah *et al.*, (Abdullah, Ameen, Sadeeq, & Zeebaree, 2021) involved in the DL-MER task, but they are limited to the brief coverage and coarse-grained introduction. This paper is the first effort on comprehensively reviewing the DL-MER literatures, more specifically, we focus on (1) summarizing the development tendency of emotion recognition from single modality to multiple modalities, from hand-crafted to deeply-learned, by analyzing the milestones; (2) providing an in-depth analysis of existing DL-MER methods involved in audio, visual and text modalities; (3) discussing several potential research directions with under-investigated issues to narrow the gap between the DL-MER methodology and practical HCI applications.

To address the above-mentioned issues, we have conducted a comprehensive literature search between 2010 and 2023 through Google Scholar and Web of Science, in terms of the main keywords: "emotion recognition", "deep learning", "audio", "speech", "visual", "text", "bimodal", "trimodal", and "multimodal". For targeted literature selection, we concentrated on those published papers in some important IEEE, Elsevier, and Springer journals and international conferences related to affective computing such as CVPR, ICCV, NIPS, AAAI, ACM MM, ACL, EMNLP, ICASSP, ICME, ACII and so on.

The rest of this paper is organized as follows. Section 2 overviews the development tendency of emotion recognition, representative deep learning techniques for DL-MER, standard DL-MER datasets, and evaluation metrics. Section 3 describes state-of-the-arts of multimodal feature extraction methods. Section 4 reviews multimodal information fusion methods. Research challenges and open issues are discussed in Section 5. Conclusions will be drawn in Section 6.

2. DL-MER Overview

2.1. The structure of general MER system

Generally, as shown in Fig. 1, building a DL-MER system requires three key steps as follows:

Step 1 - multimodal *feature extraction:* effective feature representations characterizing human emotion expression for various modalities such as audio, visual, text, *etc*., are extracted.

Step 2 - *multimodal information fusion*: different fusion strategies to integrate two or more modalities for emotion recognition are employed. Common multimodal information fusion approaches include feature-level fusion, decision-level fusion, as well as model-level fusion.

Step 3 - *emotion classifier design*: suitable classifiers are leveraged to learn the mapping relationships between the extracted feature representations and target emotions, so as to obtain target emotion labels (discrete or dimensional categories) as ultimate emotion recognition results.

Most of existing machine learning methods can be used for emotion classification or prediction. The representative classifiers include Bayesian networks, multi-layer perception (MLP), *k*-nearest neighbors (KNN), support vector machines (SVM), *etc*. (Keerthi, Shevade, Bhattacharyya, & Murthy, 2001; Rish, 2001; Windeatt, 2006). Since emotion classifier design is relatively mature, this paper focuses on recent developments of the above-mentioned Step 1 and 2.

Table 1
Comparisons between our paper and existing related surveys.

Category	Publication	Focus				Main concerns
		Feature extraction		Multimodal fusion	Multimodal Emotion Datasets	
		Hand-crafted	Deeply-learned			
Multimodal	(Sebe, Cohen, Gevers, & Huang, 2005)	√	×	√	×	Multimodal emotion recognition
	(Zeng, Pantic, Roisman, & Huang, 2008)	√	×	√	×	Audio-visual emotion
	(Wu, Lin, & Wei, 2014)	√	×	√	×	Audio-visual emotion
	(Sidney K D'mello & Jacqueline Kory, 2015)	×	×	√	×	Multimodal affect detection
	(Jiang et al., 2020)	√	×	√	Few	Multimodal information fusion
	(Zhang, Yin, Chen, & Nichele, 2020)	√	Few	Few	Few	EEG based emotion
Deep Learning	(Pandey, Shekhawat, & Prasanna, 2019)	×	Few	×	×	Deep learning, speech emotion
	(Khalil et al., 2019)	×	√	×	×	Deep learning, speech emotion
	(Lieskovská, Jakubec, Jarina, & Chmulík, 2021)	×	√	×	×	Deep learning, speech emotion
	(Jahangir, Teh, Hanif, & Mujtaba, 2021)	×	√	×	×	Deep learning, speech emotion
	(Abbaschian, Sierra-Sosa, & Elmaghraby, 2021)	×	√	×	×	Deep learning, speech emotion
	(Latif et al., 2021)	×	√	×	×	Deep learning, speech emotion
	(Islam et al., 2021)	×	√	×	×	Deep learning, EEG based emotion
	(Li & Deng, 2022)	×	√	×	×	Deep learning, facial expression
	(Canal et al., 2022)	×	√	×	×	Deep learning, facial expression
	(Yadav & Vishwakarma, 2020)	×	√	×	×	Deep learning, text emotion
	(Peng et al., 2021)	×	√	×	×	Deep learning, text emotion
DL-MER	(Latha & Priya, 2016)	×	√	×	×	Speech emotion and facial expression deep learning
	(Gu et al., 2021)	×	Few	Few	×	Video and speech emotion, deep learning
	(Sharmeen M Saleem Abdullah Abdullah et al., 2021)	×	Few	Few	×	Facial expression and speech emotion, deep learning
	Our paper	√	√	√	√	Multimodal emotion recognition, deep learning

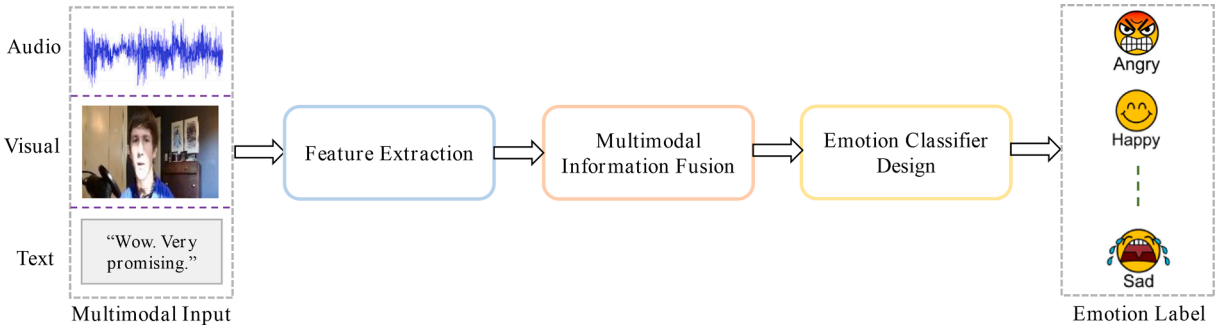


Fig. 1. The workflow of an MER system, including three main steps: (1) multimodal feature extraction, (2) multimodal information fusion, and (3) emotion classifier design.

2.2. Emotion recognition milestones

With the aid of the unflagging efforts of HCI and artificial intelligence researchers, emotion recognition has gained remarkable breakthrough in various aspects. We draw a timeline to describe some crucial milestones for emotion recognition (as shown in Fig. 2), and categorize existing methods into two main trends:

Single Modality VS Multiple Modalities: Due to the limitations of data

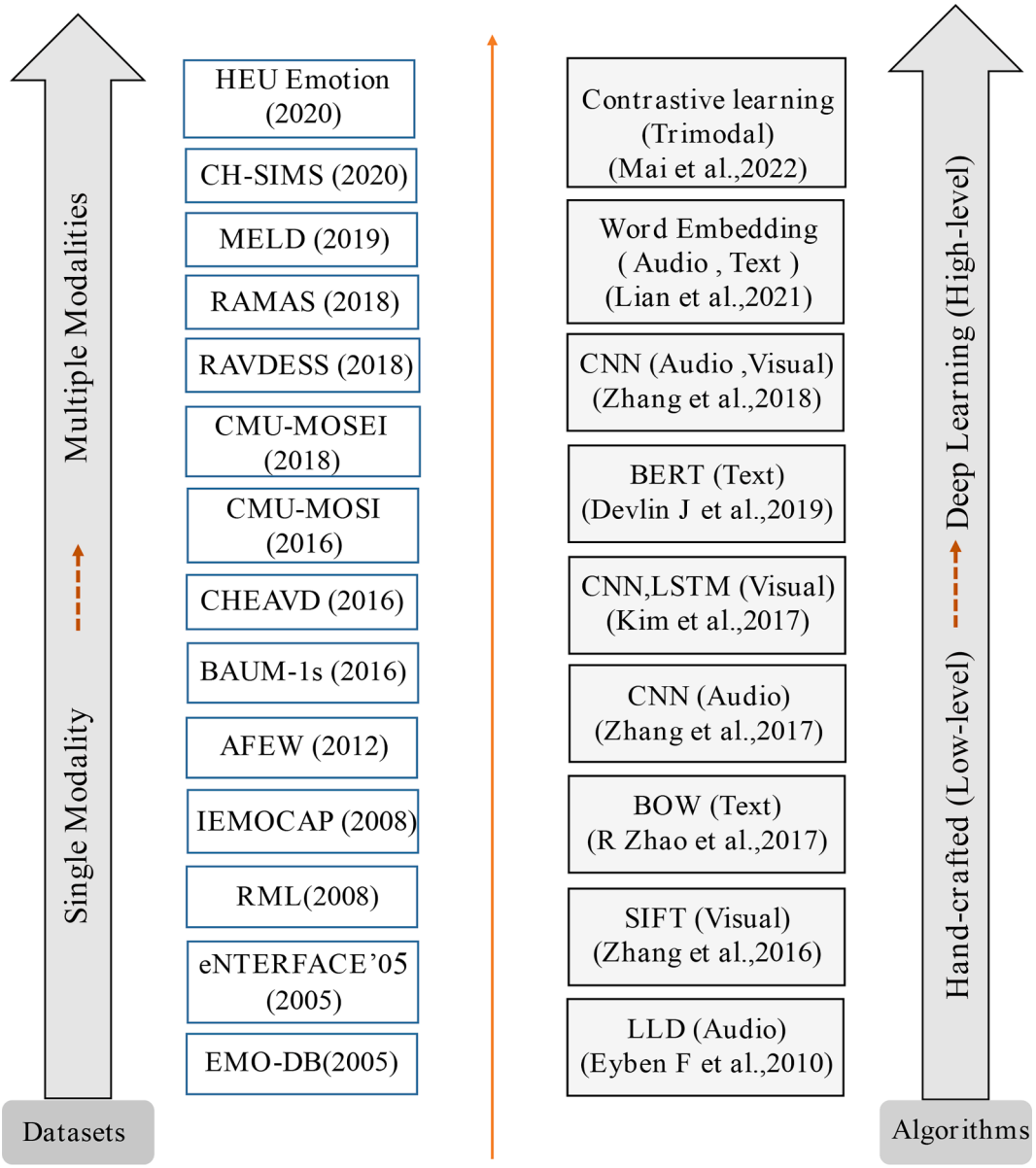


Fig. 2. The milestones of DL-MER development.

acquisition and application scenarios, early studies concentrate on single-modal emotion recognition. The conventional single-modal emotional datasets contain EMO-DB (Burkhardt, Paeschke, Rolfes, Sendlmeier, & Weiss, 2005) (audio), international affective picture system (IAPS) (Lang, 2005) (visual image), Twitter Sentiment (Go, Bhayani, & Huang, 2009) (text), etc. However, the single modality cannot accurately characterize emotion expression. To alleviate this problem, emotion recognition should consider multimodal cues, such as audio, visual, text, *etc*. To this end, various multimodal emotional datasets have been developed, such as eNTERFACE'05 (Martin, Kotsia, Macq, & Pitas, 2006), RML (Y. Wang & Guan, 2008), *etc*. The representative multimodal emotional datasets with four modalities are IEMOCAP (Busso et al., 2008), and RAMAS (Perepelkina, Kazimirova, & Konstantinova, 2018). Table 2 provides a summary of multimodal emotional datasets mentioned above.

Hand-crafted VS Deeply-learned: Feature extraction is a vital step for MER tasks. The conventional features for MER can be categorized into two groups: hand-crafted features and deeply-learned features.

Hand-crafted features have been widely applied for identifying emotion in early related works. The common audio hand-crafted features on speech emotion recognition (SER) tasks are low-level descriptors (LLD), including prosody, voice quality, spectral features, *etc*. Eyben *et al.*, (Eyben et al., 2010) developed a popular feature extraction toolkit called OpenSMILE to derive conventional audio LLD features. The conventional hand-crafted visual features on facial expression recognition (FER) tasks for static facial images are local binary pattern (LBP) (Chen, Liu, Tu, & Aragones, 2013; Zhang et al., 2012a), scale invariant feature transform (SIFT) (Zhang et al., 2016), histograms of oriented gradients (HOG) (Baltrušaitis, Mahmoud, & Robinson, 2015), Gabor wavelets (Ahsan, Jabid, & Chong, 2013), *etc*. Zhang *et al.*, (Zhang et al., 2016) extracted SIFT features for a set of marker points of each facial image. A typical hand-crafted text feature extraction method is bag-of-words (BoW) (Zhao & Mao, 2017) model.

Recently, with the adventure of deep learning approaches, deeply-learned features obtained by various deep learning algorithms have been successfully utilized for emotion classification tasks. For deep audio feature extraction, convolutional neural networks (CNNs) (LeCun, Bottou, Bengio, & Haffner, 1998) have been commonly leveraged to capture high-level speech emotional feature representations from the raw speech signals (Ahmed et al., 2023; Dai et al., 2021a; Kwon, 2021; Zhang et al., 2018; Zhang et al., 2022; Zhang et al., 2023. For deep visual feature extraction, combining CNNs with long short-term memory (LSTM) (Hochreiter & Schmidhuber, 1997) are usually implemented to learn high-level temporal-spatial feature representations (Sepas-Moghaddam, Etemad, Pereira, & Correia, 2020; Zhang, Pan, Cui, Zhao, & Liu, 2019). For deep text feature extraction, pre-trained deep learning-based word embedding methods, such as Word2vec (Mikolov et al., 2013), Glove (Pennington, Socher, & Manning, 2014), BERT (bidirectional encoder representations from Transformers) (Devlin et al., 2019), and so on, are often adopted. These representative deep learning techniques are elaborated as described below.

2.3. Representative deep learning techniques

Deep learning is regarded as an emerging area of machine learning, and has recently obtained extensive attention. Compared with traditional methods, deep learning techniques for emotion recognition have many advantages, such as the capability to automatic detection of complex structures, and extraction of high-level features from input data (Deng & Yu, 2014). In this section, several representative deep learning algorithms are briefly reviewed.

(1) Deep Belief Networks

Deep belief networks (DBNs) are considered to be a generative model, proposed by Hinton *et al.*, in 2006 (Hinton & Salakhutdinov, 2006). A DBN is a multi-layer deep structure constructed from a series of stacked restricted Boltzmann machines (RBMs) (Freund & Haussler, 1991). A RBM consists of a visible layer and a hidden layer. Every neuron in these two layers is fully connected to the neurons in another layer. Nevertheless, no connections emerge between these neurons in the same layer. DBNs' training requires two key steps: pre-training as

Table 2
Overview of multimodal emotional datasets.

Methods	Dataset	Year	Modality	Emotion label	Samples
Bimodal	eNTERFACE'05 (Martin et al., 2006)	2006	Audio, visual	anger, disgust, fear, happiness, sadness, and surprise	1,277 video clips, 42 subjects
	RML(Y. Wang & Guan, 2008)	2008	Audio, visual	anger, disgust, fear, happiness, sadness, and surprise	720 video clips, 8 subjects
	AFEW (Dhall et al., 2012)	2012	Audio, visual	anger, happiness, sadness, surprise, disgust, fear and neutral	1,426 video clips, 330 subjects
	BAUM-1 (Zhalehpour et al., 2016)	2016	Audio, visual	anger, joy, sadness, disgust, fear, surprise, boredom and contempt	1,222 video clips, 31 subjects
	CHEAVD (Ya Li et al., 2016)	2016	Audio, visual	anger, happiness, sadness, worried, anxious, surprise, disgust and neutral	140 min of video clips, 238 subjects
	RAVDESS (Livingstone, 2018)	2018	Audio, visual	calmness, happiness, sadness, anger, fear, surprise, disgust, and neutral	60 speeches, 44 songs, 24 subjects
Trimdal	CMU-MOSI (A. Zadeh et al., 2016)	2016	Audio, visual, text	positive and negative	2,199 utterances, 93 videos, 89 subjects
	CMU-MOSEI (A. B. Zadeh et al., 2018)	2018	Audio, visual text	anger, disgust, fear, happiness, sadness, and surprise	3,229 video clips with 22,676 utterances, about 1,000 subjects
	MELD(Poria et al., 2019)	2019	Audio, visual, text	anger, disgust, fear, joy, neutral, sadness, surprise	1,433 conversations, 6 leading actors
	CH-SIMS (W. Yu et al., 2020)	2020	Audio, visual, text	negative, weakly negative, neutral, weakly positive and positive	2,281 video clips
	HEU(J. Chen et al., 2021)	2021	Audio, visual, body posture	anger, bored, confused, disappointed, disgust, fear, happy, neutral, sad, surprise	1,9004 video clips, 9,951 subjects
Multimodal	IEMOCAP (Busso et al., 2008)	2008	Audio, visual, text, body posture	Category label: anger, happiness, excitement, sadness, frustration, fear, surprise, neutral and other; Dimensional labels: valence, arousal and dominance	10,039 conversations, 10 subjects
	RAMAS (Perepelkina et al., 2018)	2018	Audio, visual, body posture, physiological signals	anger, disgust, happiness, sadness, scare, surprise	7 h of video clips, 10 subjects

well as fine-tuning. Pre-training is achieved in an unsupervised manner through a layer-wise greedy learning scheme (Bengio, Lamblin, Popovici, & Larochelle, 2007). In the pre-training process, a contrastive divergence approach (Geoffrey E Hinton, 2002) is utilized for training RBMs in a DBN so as to optimize the weights and biases of the DBN model. Then, a back-propagation strategy is used to perform a fine-tuning so as to update the network parameters.

(2) Convolutional Neural Networks

Convolutional neural networks (CNNs) were originated in 1998, and developed by LeCun *et al.*, (LeCun et al., 1998). The basic structure of a CNN includes one convolutional layer, one pooling layer, and one fully-connected (FC) layer. The convolutional layer uses multiple learnable filters to convolute the entire input image, resulting in corresponding mappings of active features. The pooling layer is attached behind the convolutional layer, and aims to achieve the translational invariance by using a nonlinear down-sampling strategy in an effort to reduce the dimensionality of extracted features. Two typical pooling methods are max-pooling and average-pooling. The FC layer is usually located at the end of a CNN, and aims to activate the previous layers in order to generate final abstract feature representations. The output layer of CNNs is the Softmax operation commonly used for classification tasks. By minimizing the loss function, the stochastic gradient descent (SGD) (Bottou, 2012) strategy is usually to obtain the optimization of the CNN network in the process of training CNNs.

(3) Recurrent Neural Networks

Recurrent neural networks (RNNs) (Elman, 1990) are capable of capturing temporal information from sequence data, and thus suitable for sequence processing. RNNs employ recursive connections on the hidden states to capture historical information about sequence data. Besides, on all time steps RNNs are able to share the common network parameters. The classical back propagation trough time (BPTT) (Werbos, 1990) algorithm is adopted to train RNNs. However, RNNs are prone to suffering from gradient vanishing or exploding problems. To alleviate this problem, long short-term memory (LSTM) (Hochreiter & Schmidhuber, 1997) was proposed in 1997. In a LSTM cell unit, three different gates are included such as input gate, forget gate, and output gate. Input gate is leveraged to determine the amount of current input data stored in the memory cell unit. Forget gate is utilized to decide which information should be thrown away from the cell state. Output gate aims to control what information should be output. Based on these three special gates, LSTMs are able to model long-term dependency in sequence data.

(4) Attention Neural Networks

Attention neural networks (Niu, Zhong, & Yu, 2021) are a type of deep learning architectures equipped with the attention mechanism. The used mechanism enables the network to concentrate on particular parts of input data, thereby enhancing its performance in various tasks. The first attention neural network was attention-based RNN (Chorowski et al., 2014). More recently, the Transformer techniques leveraging a unique self-attention mechanism (Vaswani et al., 2017a) have attracted extensive interests, owing to their strong capabilities of modeling long-term dependencies. So far, a variety of Transformer-based methods, such as vision Transformer (ViT) (Khan et al., 2022), audio Transformer (Gong et al., 2022), video Transformer (Selva et al., 2023), and so on, have been developed in recent years. In addition, a number of Transformer-based methods have been successfully applied in image classification (Sun et al., 2022), object detection and segmentation (Dai, Liu, Tang, Wu, & Song, 2022; Zhao et al., 2023), speech signal processing (Dang, Chen, & Zhang, 2022), automated depression detection (Zhang et al., 2023), air quality prediction (Zhang & Zhang, 2023; Zhang, Zhang, Zhao, Chen, & Yao, 2022), *etc*. However, Transformer-related methods are rarely used in the field of multimodal emotion recognition.

It is pointed out that the advantage of deep learning is that it can effectively learn high-level features from input data for multimodal emotion classification. However, for most deep leaning methods the computational complexity is very high.

2.4. Multimodal emotion datasets

Multimodal emotion datasets refer to datasets that contain dynamic emotion changes involved in two and more emotion modalities. According to the number of involved emotion modalities, this work categorizes them into three types: bimodal, trimodal, and multimodal for better presentation, as shown in Table 2. These representative multimodal emotion datasets are described below in brief.

eNTERFACE'05 (Martin et al., 2006): It consists of 1,277 audiovisual video samples, collected by 42 subjects (8 women) from 14 countries. Each subject was asked for listening to 6 short stories, each of which evoked a specific emotion. Subjects had to respond to each situation, and two human experts were invited to judge whether those responses characterized the intended emotions in a clear-cut manner. The six specific emotions in this dataset are anger, disgust, fear, happiness, sadness, and surprise.

RML (Wang & Guan, 2008): It comprises of 720 audiovisual video samples from 8 subjects. Each video has the duration between 3 and 6 s. This dataset contains 6 basic emotions: anger, disgust, fear, happiness, sadness, and surprise. The recordings were collected by using a digital camera in a lightful background atmosphere. 8 subjects were invited to speak 6 different languages, including English, Mandarin, Urdu, Punjabi, Persian, and Italian.

AFEW (Dhall, Goecke, Lucey, & Gedeon, 2012): It includes 1,426 short video clips from 330 subjects. These video clips were annotated as one of 6 basic emotions: anger, happiness, sadness, surprise, disgust, fear and neutral. This dataset aims to capture dynamic, temporal facial expressions in close-to-real-world environments.

BAUM-1 (Zhalehpour, Onder, Akhtar, & Erdem, 2016): It is an audiovisual spontaneous dataset containing 1,222 video clips from 31 Turkish persons. This dataset is comprised of 6 basic emotions (joy, anger, sadness, disgust, fear, surprise) and other two emotions like boredom and contempt. Additionally, another four mental states (unsure, thinking, concentrating, bother) are included. To achieve spontaneous emotion expression, watching a film was adopted for emotional stimulation.

CHEAVD (Li et al., 2016): It is a Chinese spontaneous emotional audiovisual dataset, including 140 min of video clips collected from movies, TV plays as well as talk shows. This dataset has 238 subjects (125 male, 113 female), covering from children to the elderly. 4 native speakers were asked to annotate each video clip. 8 major emotions are included in this dataset, such as anger, happiness, sadness, worried, anxious, surprise, disgust and neutral.

RAVDESS (Livingstone, 2018): It is recorded by 24 professional actors and contains 60 speeches as well as 44 songs with 8 emotion categories (neutral, calmness, happiness, sadness, anger, fear, disgust, surprise). The recordings of each actor are available in three formats: audio-visual, visual and audio. The recording was collected in a professional studio with only the actors and green screen visible in the footage.

CMU-MOSI (Zadeh, Zellers, Pincus, & Morency, 2016): It contains 2,199 opinionated utterances and 93 videos from 89 subjects. These videos cover a wide range of topics related to movie, book, and product. It takes the first 62 video samples as the training data, and the other 31 video samples as the testing data. There are 1,447 utterances and 752 utterances in training and testing sets, respectively. Video data were scraped from the online YouTube and segmented into different utterances. 5 annotators were asked to label each utterance with a score between +3 (strong positive) and −3 (strong negative), and the mean scores of these five annotations were taken as the final emotion polarity.

CMU-MOSEI (Zadeh, Liang, Poria, Cambria, & Morency, 2018): It includes 3,229 videos from over 1,000 online YouTube speakers with 6 basic emotions: anger, disgust, fear, happiness, sadness, as well as surprise. It is annotated at the utterance level with a total of 23,259

samples. The samples consist of three modes: audio data sampled at 44.1 kHz, text transcription, and image frames sampled from video at 30 Hz. This dataset is gender-balanced, and all sentences are randomly selected from a variety of topics and monologue videos, which are transcribed and marked with correct punctuation.

MELD (Poria et al., 2019): It contains 13,000 sentences from 1,433 dialogues in the TV series "Friends", and each dialogue includes more than two speakers. Since this dataset was only collected from one TV series, the number of subjects was small, and 84% of the episodes were achieved by 6 leading actors. Every sentence in the conversation was tagged with one of these 7 basic emotions: anger, disgust, sadness, joy, neutral, surprise, as well as fear. Additionally, it also has three kinds of emotional labeling such as positive, negative and neutral for each utterance.

CH-SIMS (Yu et al., 2020): It consists of 2,281 refined wild video clips with multimodal and independent unimodal annotations. This dataset only considers Mandarin Chinese and is cautious about the choice of accent material. Each clip should be more than 1 s long and no more than 10 s long. Each clip contains 15 words and has an average length of 3.67 s. They were labeled by multiple human annotators according to the average of five emotional scores, such as negative, weakly negative, neutral, weakly positive, as well as positive.

HEU (Chen et al., 2021): It is composed of 19,004 video clips from 9,951 subjects from different languages such as Chinese, Americans, Thais and Koreans. In terms of data source, it is divided into two subsets: HEU-part1 and HEU-part2. More specially, HEU-part1 consists of 16,569 video clips from 8,984 subjects available from Tumblr, Google and Giphy, and contains 10 emotions (anger, bored, confused, disappointed, disgust, fear, happy, neutral, sad, surprise) from two different modalities like facial expressions and body postures. HEU-part2 consists of 2,435 video clips from 967 subjects available from movies, TV dramas, and variety shows. HEU-part2 contains 10 similar emotions from three different modalities like facial expressions, body postures, as well as speech signals.

IEMOCAP (Busso et al., 2008): It is collected by the Sail lab at the University of Southern California and includes behavior of 10 subjects in conversation, including video, speech signals, facial motions, and text transcriptions. In total, it contains 10,039 conversations with a mean duration of 4.5 s and an average word count of 11.4. Participants perform improvised or scripted scenes. It has been marked as neutral, happy, sad, anger, surprise, fear, disgust, frustration, excitement and other categorized labels and dimensional labels like valence, arousal, and dominance by multiple annotators.

RAMAS (Perepelkina et al., 2018): It is a Russian multimodal sentiment database and consists of about 7 h of high-quality close-up video recordings, capturing a variety of emotional modalities such as speech signals, body postures, panoramic video, as well as physiological signals. 10 semi-professional actors with 18–28 years old (5 female and 5 male) from Russian were invited to participate in the data collection. Semi-professional actors expressed 6 basic emotions, including anger, disgust, happiness, sadness, scare, as well as surprise.

Highlights in this section: multimodal datasets have multiple advantages, such as rich emotional representations and higher accuracy compared to single modal datasets; but it also faces many challenges, for example, it is difficult to collect large-scale annotated multimodal emotional datasets since it requires a lot of time-consuming and labor-intensive costs.

Feature Extraction for MER.

Feature extraction is the first vital step in a MER system. This section focuses on feature extraction methods related to audio/speech, visual, and text emotions that are the most common three modalities. Both hand-crafted and deeply-learned features for each modality will be described in detail. Table 3–5 separately shows a summary of feature extraction methods for the single modality.

Table 3
A summary of audio feature extraction methods.

Type of features	Feature extraction methods	Publications
Hand-crafted Speech Features	Pitch, energy and audible duration	(Yacoub et al., 2003)
	MFCCs	(Schmitt et al., 2016)
	Prosody, voice quality and spectral features	(Luengo et al., 2010)
	Prosody and voice quality features	(Zhang & Zhao, 2013)
Deeply-learned Speech Features	Hybrid DBN-HMM	(Le & Provost, 2013)
	CNNs	(Mao et al., 2014)
	CNNs	(Zhang, S. Zhang, T. Huang, & W. Gao, 2018)
	1D CNN, 2D CNN, and 3D CNN	(Zhang, X. Tao et al., 2021)
	Multi-scale convolutional LSTM	(Zhang, Zhao et al., 2019)
	CNNs	(Ottl et al., 2020)

Table 4
A summary of visual feature extraction methods.

Input type	Type of features	Feature extraction	Publications
Static face images	Hand-crafted Visual Features	LBP	(Shan et al., 2009)
		SIFT, HOG, and LBP	(Hu et al., 2008)
		Raw pixels, Gabor wavelets, LBPs	(Zhang et al., 2012b)
		MB-LBPUH	(Xia et al., 2020)
	Deeply-learned Visual Features	DBNs	(Zhao et al., 2015)
		1D CNNs	(Zhang et al., 2016)
		Four-stage CNN networks	(Yolcu et al., 2019)
		MDSTFN	(Sun et al., 2019)
		Attention-based Bi-LSTM	(Sepas-Moghaddam et al., 2020)
		CNNs with attention mechanism	(Li et al., 2019)
Dynamic video sequences	Hand-crafted Visual Features	Optical flow	(Yeasin et al., 2004)
		FHOFO	(Happy & Routray, 2017)
		VLBP	(Zhao & Pietikainen, 2007)
		Spatio-temporal descriptor	(Fan & Tjahjadi, 2015)
		Feature point movements	(Xu et al., 2019)
	Deeply-learned Visual Features	Hybrid deep learning	(Zhang, Pan et al., 2019)
		Joint fine-tuning CNNs	(Jung, Lee, Yim, Park, & Kim, 2015)
		CNN-RNN, 3D-CNN	(Fan, Lu, Li, & Liu, 2016)
		CNNs, LSTMs	(Kim, Baddar, Jang, & Ro, 2017)
		STC-NLSTM	(Yu, Liu, Liu, & Deng, 2018)
		GANs	(Guo, Zhao, Zhang, & Pan, 2022)
		AC-GAN	(Joseph Raj, & Gopi, 2021)
		Former-DFER	(Zhao & Liu, 2021)

2.5. Speech feature extraction

(1) Hand-crafted Speech Features

The conventional emotional speech features in early works related to speech emotion recognition (SER) were hand-crafted low-level descriptors (LLD), including prosody features, voice quality features, spectral features, and so on (Akçay & Oğuz, 2020; Swain, Routray, & Kabisatpathy, 2018). The representative prosody features is composed of pitch, energy, amplitude, duration, *etc*. (Ten Bosch, 2003). The typical

Table 5
A summary of text feature extraction methods.

Type of features	Feature extraction	Publications
Hand-crafted Text Features	BoW	(Sebastiani, 2002; Soumya George & Joseph, 2014)
	BoW	(Colnerič & Demšar, 2018)
	LDA	(Blei et al., 2003)
	A joint emotion-topic model	(Bao et al., 2012)
	LSA	(Deerwester et al., 1990)
	Bi-words occurrence	(Inrak & Sinthupinyo, 2010)
Deeply-learned Text Features	Word2vec, GloVe	(Mikolov et al., 2013), (Pennington et al., 2014)
	CoVe, ELMo	(Tan & Celis, 2019), (Peters et al., 2018)
	Transformer	(Vaswani et al., 2017b), (Chung & Glass, 2020), (Radford et al., 2019), (Brown et al., 2020)
	BERT	(Devlin et al., 2019)
	Transformer-XL, XLNet	(Dai et al., 2019), (Yang et al., 2019)
	BERT combined with a Bi-LSTM	(Cai et al., 2020)
	Emo2Vec	(Xu, Madotto, Wu, Park, & Fung, 2018)
	Emoji2vec	(Eisner, Rocktäschel, Augenstein, Bošnjak, & Riedel, 2016)
	DeepMoji	(Felbo, Mislove, Søgaard, Rahwan, & Lehmann, 2017)
	ReferEmo, DeepMoji	(Esperanca & Luo, 2022)
	Learning domain sensitivity and emotion perception embedding	(Shi, Fu, Bing, & Lam, 2018)
	Multi-task CNNs	(Zhang et al., 2018)
	Stacked ensemble of CNN, LSTM, and GRU	(Akhtar, Ekbal, & Cambria, 2020)
	ConvLexLSTM	(Khanpour & Caragea, 2018)

voice quality features consist of formant, spectral energy distribution, glottal features, *etc*. (Sundberg, Patel, Bjorkner, & Scherer, 2011). The common spectral features contain Mel-frequency cepstral coefficients (MFCCs), linear predictive coding (LPC), perceptual linear production (PLP), and so on (Y. Sun, Wen, & Wang, 2015).

Yacoub *et al.*, (Yacoub, Simske, Lin, & Burns, 2003) extracted prosody features like pitch, energy and audible duration, and compared the performance of neural networks, binary SVM, K-nearest neighbors (KNN), and decision trees on SER tasks. They found that neural networks outperformed other classifiers on SER tasks. Schmitt *et al.*, (Schmitt, Ringeval, & Schuller, 2016) employed a bag-of-audio-words (BoAW) method created by MFCCs as feature representations, and then leveraged a SVM-based regression method to implement time-continuous prediction of emotional arousal and valence values. Luengo *et al.*, (Luengo, Navas, & Hernáez, 2010) extracted acoustic parameters related to prosody, voice quality and spectral features from speech signals. They analyzed individual parameters and investigated the combination of different parameters to compare the different performance of these features on SER tasks. Zhang *et al.*, (Zhang & Zhao, 2013) adopted prosody and voice quality features, and then performed nonlinear dimensionality reduction by using modified supervised locally linear embedding algorithm (MSLLE) to achieve low-dimensional discriminating embedded feature representations for SER.

In recent years, some typical feature sets equipped with thousands of high-level statistical parameters of LLDs, including Interspeech-2010 (Schuller et al., 2010), ComParE-2013 (Schuller et al., 2013), GeMAPS and its extension called eGeMAPS (Eyben et al., 2016), have also conventionally employed for SER. Chen *et al.*, (Chen et al., 2020) provided a two-layer fuzzy multiple random forest (TLFMRF) model for SER. They initially extracted personalized and non-personalized features consisting of ComParE-2013 features with the aid of OpenSMILE (Eyben et al., 2010). Then, a fuzzy C-means clustering scheme was utilized to divide high-dimensional speech features into different subclasses. Finally, multiple random forests were utilized to distinguish emotion categories of selected speech features.

(2) Deeply-learned Speech Features

In recent years, various deep learning algorithms have been adopted for deep speech feature extraction on SER tasks. These representative deep learning methods for speech feature extraction include DBNs, CNNs, RNNs/LSTMs, *etc*, as described below.

Le *et al.*, (Le & Provost, 2013) presented a hybrid DBN-HMM model for SER. They adopted a DBN with 5 hidden layers for modeling complex and non-linear connections between these extracted MFCC features. Then, they modeled each emotion by using a left-to-right HMM. Mao *et al.*, (Mao, Dong, Huang, & Zhan, 2014) adopted CNNs to extract emotion-salient feature representations for SER in two stages. First, local invariant features were learned by a sparse auto-encoder. Second, local invariant features were used to perform salient discriminating feature analysis so as to capture emotion-salient and discriminating feature representations for SER. Zhang *et al.*, (Zhang, Zhang, Huang, & Gao, 2018) utilized CNNs to extract high-level feature representations from every divided speech segment from an utterance, and then employed a discriminant temporal pyramid matching strategy to aggregate segment-level feature representations for producing utterance-level features on SER tasks. They fine-tuned the pre-trained AlexNet model to capture high-level feature representations on every divided speech segment. To this end, they proposed to create three channels of log Mel-spectrograms similar to a RGB image as an input of CNNs. Zhang *et al.*, (Zhang, Tao, Chuang, & Zhao, 2021) employed multi-CNNs, such as 1D CNN, 2D CNN, and 3D CNN, to separately extract deep multimodal segment-level feature representations for SER. Then, an average-pooling was used to yield utterance-level SER results, followed by a score-level fusion scheme for integrating these utterance-level classification results. They showed that the learned deep multimodal speech representations were complementary to each other. Zhang *et al.*, (Zhang, Zhao, & Tian, 2019) presented a multi-scale deep convolutional LSTM model for SER, which was inspired by different impacts of varying lengths of speech spectrograms on SER tasks. Initially, they adopted CNNs to extract deep segment-level features, followed by LSTMs to learn the temporal dependencies among all divided speech segments. Then, different SER results, achieved by integrating CNNs with LSTMs at various lengths of segment-level spectrograms, were combined with a score-level fusion method to conduct final SER tasks. Ottl *et al.*, (Ottl, Amiriparian, Gerczuk, Karas, & Schuller, 2020) leveraged a deep spectrum system (Amiriparian et al., 2017) to achieve deep image-based features from the audio contents of EmotiW-2020. Then, several different CNN architectures pre-trained on ImageNet data, such as AlexNet, VGG16, and three modified DenseNets, were adopted for speech feature extraction tasks. These CNNs-based results were combined by early and late fusion. They also used OpenSMILE (Eyben et al., 2010) to extract ComParE-2013 features to compare its performance with deep spectrum features.

Highlights in this section: (1) Leveraging the professional OpenSMILE (Eyben et al., 2010) tool to obtain LLD features for SER, has become very popular for hand-crafted speech feature extraction. (2) Among typical deep learning methods, CNNs have been more prevalent than other deep models, since CNNs are able to capture high-level speech feature representations from the original speech signals. (3) Combining LSTMs with CNNs has been currently fashionable due to its good spatio-temporal feature learning ability. (4) Although deeply-learned speech representations usually outperform hand-crafted speech features on SER tasks, both of them may have their own advantages and disadvantages. Hand-crafted speech features belong to low-level features, which are simple for extraction. By contrast, deeply-learned speech features are high-level, and thus can more effectively characterize speech emotion expression. Nevertheless, the computational cost for deeply-learned speech features is high. In this case, how to effectively fuse these two distinct features for improving SER performance is a meaningful research direction.

2.6. Visual feature extraction

According to the processed type of input data for visual emotion classification, namely facial expression recognition (FER), we divide prior works related to visual feature extraction in literatures into two groups: static face images as well as dynamic video sequences for hand-crafted or deeply-learned feature extraction.

(1) Hand-crafted Visual Features

Static face images: static face images refer to still facial images without containing temporal information (Wu et al., 2022; Wu et al., 2020; Wu et al., 2021). After implementing a series of pre-processing, such as face localization, alignment, normalization, *etc*., the geometry and appearance features are usually extracted to characterize facial expressions. The early-used hand-crafted visual feature extraction methods for static face images are local binary patterns (LBP) (Chen et al., 2013; Zhao & Zhang, 2016), scale invariant feature transform (SIFT) (Chu, De la Torre, & Cohn, 2016), histograms of oriented gradients (HOG) (Baltrušaitis et al., 2015), Gabor wavelet representations (Ahsan et al., 2013), *etc*.

Shan *et al*., (Shan, Gong, & McOwan, 2009) adopted a LBP to empirically evaluate facial feature representations for FER. They further formulated a boosted-LBP to achieve the most discriminative LBP features, and reported the highest classification accuracy based on SVM and boosted-LBP features on FER tasks. Hu *et al*., (Hu et al., 2008) presented a performance comparison on multi-view facial expression recognition tasks. They leveraged three different local patch descriptors, such as SIFT, HOG, and LBP, to capture facial feature representations, as an input of a nearest-neighbor indexing approach for identifying facial expression. Zhang *et al*., (Zhang et al., 2012b) provided a robust FER method based on a sparse representation classifier. Three kinds of facial features, including raw pixels, Gabor wavelet representations as well as LBPs, were derived to investigate the performance of the proposed approach in identifying clean and occluded facial expression tasks. Wang *et al*., (Xia, Wang, Song, Chen, & Li, 2020) developed a facial expression representation of weighted-fusion features for FER. To characterize the holistic structural features, they initially employed an operator called multi-scale block local binary pattern uniform histogram (MB-LBPUH) for filtering facial images. Then, they concatenated a MB-LBPUH and a HOG into a new feature representation to represent facial expressions.

Dynamic video sequences: facial expressions involve in a dynamic process, thereby containing dynamic information such as facial muscle movements and facial shape changes. The dynamic information can be used to effectively represent facial expressions. Therefore, it is crucial to model such dynamic information so as to identify facial expressions from video sequences. To this end, typical hand-crafted visual features for dynamic video expression sequences include optical flow (Yeasin et al., 2004), local binary patterns on three orthogonal planes (LBP-TOP) (Zhao & Pietikainen, 2007), feature point movements (Xu et al., 2019), and so on.

In video sequences, optical flow features have been widely employed to characterize facial motions for facial expression representations by means of calculating the geometric displacement of facial landmarks between two adjacent frames (Yeasin et al., 2004). Happy *et al*., (Happy & Routray, 2017) presented a fuzzy histogram of optical flow orientation (FHOFO) for identifying micro-expressions. They investigated the temporal features on the basis of facial micro-movements. The proposed FHOFO features were adopted to construct appropriate angular histograms from the extracted optical flow vector orientations so as to capture the temporal patterns for micro-expression recognition. Zhao *et al*., (Zhao & Pietikainen, 2007) developed dynamic textures for facial image analysis. They used a variant of the LBP operator called volume local binary patterns (VLBP) to model textures combining motion and appearance. Then, for computation simplicity, only the co-occurrences of LBP-TOP were taken into account. Moreover, a block-based scheme was presented to process special dynamic events like facial expressions. Fan *et al*., (Fan & Tjahjadi, 2015) presented a spatio-temporal structure on the basis of gradient histograms as well as optical flow for FER. Specially, they extended spatial pyramid gradient histograms (Bosch, Zisserman, & Munoz, 2007) to a spatio-temporal structure for producing three dimensional facial features. Then, they adopted a dense optical flow to integrate them so as to construct a spatio-temporal descriptor related to facial expressions in video sequences. Finally, a multi-class SVM was adopted to distinguish facial expressions. Yi *et al*., (Xu et al., 2019) developed a FER framework based on feature point movements as well as feature block texture variations. First, they selected the most representative 24 marked feature points obtained by an active appearance model (AAM). Second, facial expression sequences were intercepted from facial videos by means of determining two crucial frames associated with the minimum and maximum emotion intensities. Third, a trend curve representing the changes between any two feature points was achieved. Finally, a set of calculated slopes were combined with the developed feature block texture variations to form final expression features for FER.

(2) Deeply-learned Visual Features

Static face image: For deeply-learned visual feature extraction of static face image, various deep models, such as DBNs, self-built CNN models from scratch or fine-tuned several representative pre-trained CNN models, like AlexNet (Krizhevsky, Sutskever, & Hinton, 2012), VGGNet (Simonyan & Zisserman, 2014), GoogleNet (Szegedy et al., 2015), ResNet (He et al., 2016), have been widely used for FER.

Zhao *et al*., (Zhao, Shi, & Zhang, 2015) provided a FER method with DBNs. They combined DBNs with a MLP for FER. In particular, the proposed method integrated the unsupervised feature learning capability of DBNs with the classification capability of MLP classifier on FER tasks. Zhang *et al*., (Zhang et al., 2016) developed a deep learning-driven feature learning algorithm for multi-view FER tasks. They first extracted SIFT features related to a set of marker points for every facial image. Next, a feature matrix composed of the obtained SIFT features was then fed into a well-designed deep learning model consisting of several 1D convolutional operations to learn discriminative features for facial expression classification. Yolcu *et al*., (Yolcu et al., 2019) provided an automated FER system based on a deep learning architecture with four-stage CNN networks. The first three networks were used to segment essential facial components for FER, thereby producing an icon facial image. The fourth network employed original facial images as well as icon facial images to recognize facial expressions. Sun *et al*., (Sun, Li, Huan, Liu, & Han, 2019) used optical flow to capture temporal feature representations from static images, and developed a multi-channel deep spatio-temporal feature fusion neural network (MDSTFN) model for deep spatio-temporal feature extraction and fusion on FER tasks. Each channel in this model was fine-tuned by using a pre-trained CNN model like GoogleNet. Sepas-Moghaddam *et al*., (Sepas-Moghaddam et al., 2020) provided a FER method based on light field images with deep attention-based Bi-LSTM. They first extracted spatial features by using a VGG16 network. Then, a Bi-LSTM was utilized to learn spatio-angular features from the viewpoint feature sequences. In addition, the most important spatial-angular features were selectively focused by an attention mechanism. Finally, a fusion method was adopted to achieve FER results. Li *et al*., (Li, Zeng, Shan, & Chen, 2019) developed an occlusion-aware FER method via CNNs with attention mechanism (ACNN) that aims at perceiving occluded regions of facial images and concentrating on the most discriminating un-occluded regions. The proposed ACNN aimed to combine various feature representations from facial regions of interest (ROIs). Each feature representation was weighed by using an adaptive weight strategy in a gate unit.

Dynamic video expression sequence: For deep spatio-temporal visual feature extraction from video sequences, the popular approaches contain CNN-RNN/LSTM, 3D-CNN, and so on.

Zhang *et al*., (Zhang, Pan et al., 2019) provided a video-based FER method by using a hybrid deep learning strategy. This method adopted two independent CNNs for capturing high-level spatio-temporal feature

representations from the divided video fragments. One is a spatial CNN for handling static facial images. The other is a temporal CNN for handling optical flow images. Then, the achieved segment-level spatial and temporal features were combined by using a deep fusion network consisting of a DBN model for jointly learning discriminant spatio-temporal representations. Finally, a linear SVM was used for FER. Zhang *et al.*, (Jung et al., 2015) proposed a joint fine-tuning method in deep neural networks for FER in which two independent deep neural networks were jointly trained. One is a deep CNN network for obtaining temporal appearance feature representations from image sequences. The other is a linear FC network for extracting temporal geometry feature representations based on facial landmark points. Then, these two deep neural networks were integrated by using several linear FC layers so as to promote FER performance. Fan *et al.* (Fan et al., 2016) developed a video-based FER method by using CNN-RNN and 3D-CNN hybrid networks. They combined RNNs and 3D-CNN in a late fusion (LF) manner. The used RNNs and 3D-CNN were used to capture appearance and motion feature representations of video sequences in different fashions. In particular, RNNs captured appearance feature representations learned by CNNs over each video frame and encoded motion later, while 3D-CNN simultaneously modeled appearance and motion of video sequences. Kim *et al.*, (Kim et al., 2017) developed a robust spatio-temporal feature extraction approach integrating CNNs and LSTMs for FER to alleviate the problem of expression intensity differentiation. This method employed representative expression-states in facial sequences in spite of expression intensity differentiation, and encoded facial features in two parts. First, a CNN was used to capture spatial characteristics of typical expression-state frames. Second, temporal characteristics of spatial feature representations of facial expressions obtained with CNNs, were captured with a LSTM. Yu *et al.*, (Yu et al., 2018) presented an end-to-end spatio-temporal convolutional feature extraction approach equipped with a nested LSTM (STC-NLSTM) in an effort to jointly capture multi-level appearance feature representations as well as temporal dynamical information of facial expressions. Specially, a 3D-CNN was utilized to encode spatio-temporal convolutional characteristics from image sequences. The dynamical characteristics of facial expressions were learned by a nested LSTM consisting of two sub-LSTMs, namely temporal-LSTM (T-LSTM) and convolutional-LSTM (C-LSTM). Among them, T-LSTM aimed to capture the temporal dynamical characteristics of spatio-temporal convolutional features, whereas C-LSTM aimed to fuse all the outputs of T-LSTM in purpose of encoding multi-level appearance feature representations embedded in intermediate layers of the used networks.

More recently, several advanced deep learning methods, such as generative adversarial networks (GANs) (Goodfellow et al., 2014; Saxena & Cao, 2021) and Transformer (Vaswani et al., 2017a), have drawn extensive attention for dynamic FER in video sequences. GANs aim to learn deep feature representations through a competitive process involving a pair of networks like a generator and a discriminator. Transformer aims to encode input data for producing powerful features by using the attention mechanism. Some examples involved in GANs and Transformer for FER are described below.

Guo *et al.*, (Guo et al., 2022) provided a FER method of learning inter-class optical flow difference on the basis of GANs. First, they adopted a GAN for producing inter-class optical flow facial images in terms of the variations between static fully expressive facial images and neutral facial images. Second, they employed four-channel CNNs so as to separately capture high-level optical flow feature representations from the obtained inter-class optical flow facial images, as well as static appearance feature representations from fully expressive facial images. Final FER tasks were conducted by using a decision-level fusion scheme. Dharanya *et al.*, (V et al., 2021) presented a FER method by using person-wise regeneration of expressions based on an auxiliary classifier generative adversarial network (AC-GAN) model. The proposed AC-GAN aimed to regenerate several typical expressions, such as angry, disgust, fear, joy, neutral, sad, surprise, *etc.*, from input face images and identified its expressions. To alleviate the subject-dependence problem, they trained the AC-GAN model in a person-wise way and generated all the typical expressions for a person and allowed the discriminator to distinguish the expressions. The generator of the AC-GAN model adopted a U-Net (Ronneberger, Fischer, & Brox, 2015) architecture, and the discriminator employed a Capsule network (Hinton, Sabour, & Frosst, 2018) for improving feature extraction. Zhao *et al.*, (Zhao & Liu, 2021) developed a dynamic FER transformer (Former-DFER) to perform FER tasks in the wild scene. In particular, the developed method comprised of a convolutional spatial transformer as well as a temporal transformer. The used spatial transformer contained several convolutional blocks and spatial encoder blocks so as to capture robust spatial facial feature representations related to occlusion and pose. The used temporal transformer included several temporal encoder blocks, thereby allowing it to encode contextual temporal facial feature representations.

Highlights in this section: (1) For static facial images, the popular hand-crafted visual feature extraction methods contain LBP, HOG, SIFT, Gabor wavelet representations and so on, whereas the typical deeply-learned visual feature extraction methods include DBNs, self-built CNN models or fine-tuned CNN models. (2) For dynamic video expression sequences, it is necessary to capture spatio-temporal characteristics of video sequences in an effort to effectively characterize the useful properties related to facial expressions. To this end, the prevalent hand-crafted visual feature extraction methods for video sequences are optical flow, LBP-TOP, feature point movements and so on. For deeply-learned spatio-temporal visual feature extraction for video sequences, the representative CNN-RNN/LSTM and 3D-CNN are usually employed, since these deeply-learned techniques can learn high-level visual feature representations. However, the high computational cost of deeply-learned visual features is a challenge that cannot be ignored. (3) Recently-emerged GANs and Transformer techniques have been an important guiding direction for dynamic FER in video sequences.

2.7. Text feature extraction

(1) Hand-crafted Text Features

A conventional method of extracting hand-crafted emotional features from texts is bag-of-words (BoW) (Sebastiani, 2002; Soumya George & Joseph, 2014), which is leveraged for learning a vector representation of text documents. Colnerič *et al.*, (Colnerič & Demšar, 2018) investigated two methods of transforming raw texts into BoW models for emotion recognition on twitter. One is the vanilla BoW model without any normalization of tokens. The other is the normalized BoW model which aims to reduce the feature dimensionality. Although the BoW model is simple and fashionable, it still suffers from the problem of high-dimensional sparsity and missing inter-word relationships. To alleviate this problem of the BoW model, alternative improved methods are latent Dirichlet allocation (LDA) (Blei, Ng, & Jordan, 2003), and latent semantic analysis (LSA) (Deerwester, Dumais, Landauer, Furnas, & Harshman, 1990), as listed below.

LDA (Blei et al., 2003), as a popular topic modeling method, aims to extract useful topics in various documents on the basis of the procedural probability distribution. Bao *et al.*, (Bao et al., 2012) explored the relationships between social emotions as well as affective terms from text contents automatically. In particular, they developed a joint emotion-topic model by means of enhancing LDA with a newly-added layer for emotion modeling. The developed method initially produced a set of latent emotional topics, and then generated emotional terms from every topic.

LSA (Deerwester et al., 1990) used singular value decomposition (SVD) to transformer the raw BoW feature representations into low-dimension vectors. Inrak *et al.*, (Inrak & Sinthupinyo, 2010) utilized bi-words occurrence based on LSA to identify emotions hidden in short Thai texts. More specifically, they leveraged LSA to recognize Thai texts into 6 basic emotions like anger, disgust, fear, happiness, sadness, as well as surprise. They compared the recognition results of two used

models to extract textural features. The first model recognized texts by using the single word, whereas the second model classified texts by means of integrating the single word with bi-words.

(2) Deeply-learned Text Features

For deeply-learned text feature extraction, we initially describe word embedding techniques, which aim to map input sequences to a continuous feature vector and produce the underlying input representations. Next, we introduce several deep neural network architectures for text feature extraction.

As far as word embedding is concerned, according to the diverse highlighting of the encoded information, word embedding approaches can be grouped into two categories: typical word embedding and emotional word embedding (Deng & Ren, 2021). The former aims to learn continuous word embedding by means of capturing universal semantic and contextual information, whereas the latter aims to encode emotion-related information into word embedding.

Typical word embedding: The early-used word embedding methods, such as Word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014), aim at capturing fine-grained grammar and semantic regularities. For training these word embedding methods, a large amount of unlabeled data with syntactic contexts is usually needed. Nevertheless, these word embedding models suppose that a word is represented by a unique vector, thereby ignoring the effect of different contextual information.

In recent years, motivated by the powerful transfer learning ability of pre-trained CNN models in computer vision, a number of pre-trained language models in the field of NLP have been employed for text emotion recognition. The representative pre-trained language models contain CoVe (contextualized word vectors) (Tan & Celis, 2019), and ELMo (embedding from language models) (Peters et al., 2018). Then, several newly-emerged pre-trained language model based on Transformer (Vaswani et al., 2017b), such as generative pre-training (GPT) (Chung & Glass, 2020), and its variants called GPT-2 (Radford et al., 2019) and GPT-3 (Brown et al., 2020), have been developed. More recently, a BERT (Devlin et al., 2019) model has achieved remarkable performance in many NLP tasks. BERT is a deep multi-layer Transformer network architecture and trained with tremendous unlabeled text data in an unsupervised learning manner for learning useful linguistic knowledge. Subsequently, several advanced pre-trained models, such as Transformer-XL (Dai et al., 2019), and XLNet (Yang et al., 2019), have also developed. Cai *et al.*, (Cai et al., 2020) combined a BERT with a Bi-LSTM to model and forecast the sentiment orientation of users' statements in energy market. More examples about BERT-based approaches for text emotion recognition can be found in a recent survey (Acheampong, Nunoo-Mensah, & Chen, 2021).

Emotional word embedding: Recently, emotional word embedding approaches have also exhibited promising performance on some affective computing tasks like emotion classification as well as emotional intensity prediction.

To incorporate emotional information into word representations, Xu *et al.*, (Xu et al., 2018) proposed a model called Emo2Vec to encode the emotion semantic information into a fixed-size word-level feature vector. Emo2Vec was pre-trained for 6 different emotion-related tasks by means of using multi-task learning methods. Eisner *et al.*, (Eisner et al., 2016) developed a model named Emoji2vec for directly mapping emojis into a continuous feature vector. Emoji2vec was a pre-trained embedding for all Unicode emoji. Felbo *et al.*, (Felbo et al., 2017) presented a model called DeepMoji to produce emotional feature representations. DeepMoji was pre-trained on 1.2 billion Twitter data by using a two-layer attention Bi-LSTM. Esperanca *et al.*, (Esperanca & Luo, 2022) proposed a model called ReferEmo that took emojis as textual inputs and utilized DeepMoji to produce emotional feature vectors used as reference while integrating different modalities of text encoding. Shi *et al.*, (Shi et al., 2018) developed a new method of learning domain sensitivity and emotion perception embedding, which captured both emotional semantic information and domain-sensitive information of a single word, thereby facilitating sentiment classification at the sentence and vocabulary level.

Deep neural networks: So far, various deep neural network models, such as CNNs, RNNs, LSTMs, GRUs, gated RNNs (GRNNs) and so on, have been employed for extracting useful sequence information as well as semantic information for text emotion classification.

Zhang *et al.*, (Zhang et al., 2018) presented a multi-task CNN based on emotion distribution learning for text emotion recognition. Akhtar *et al.*, (Akhtar et al., 2020) developed a stacked ensemble approach to predict emotional intensity by means of integrating three deep learning models like CNN, LSTM, and GRU, respectively, and one conventional supervised model with SVR. Abdul-Mageed *et al.*, (Abdul-Mageed & Ungar, 2017) provided an emotion detection method based on distant supervision and GRNN, and obtained promising performance on recognizing 24 fine-grained emotions. Khanpour *et al.*, (Khanpour & Caragea, 2018) developed an emotion detection method in online health communities based on a model called ConvLexLSTM combining CNNs and LSTMs to yield the final output via the Softmax mechanism. More examples about deep neural networks for sentiment analysis can also be found in recent reviews (Nassif, Elnagar, Shahin, & Henno, 2021).

Highlights in this section: (1) The popular hand-drafted text feature extraction methods contain BoW, LDA, LSA, *etc*. Nevertheless, these hand-crafted models fail to capture high-level semantic meanings behind text data. (2) For deeply-learned text feature extraction, some pre-trained deep learning-oriented typical word embedding models are prevalent, such as Word2vec, Glove, BERT, *etc*. Additionally, the common pre-trained emotional word embedding models include Emo2Vec, Emoji2vec, DeepMoji, ReferEmo, *etc*. Besides, deep neural network approaches, such as CNNs, RNNs, and LSTMs, have become the most common methods for text emotion recognition. (3) Various deep learning models have been fashionable for text feature extraction, since deeply-learned text features methods have the advantage of capturing high-level semantic meanings from text data. It is thus very valuable to pay more attention to exploring advanced deep learning models for text emotion classification in future, although deep text feature learning methods have the shortcomings of high computational complexity and over-reliance on existing data.

3. Multimodal information fusion for MER

Although prior works mentioned above have achieved promising performance on single-modal emotion recognition tasks, in recent years tremendous works have shown that MER usually performs better than single-modal emotion recognition (Sharmeen M. Saleem Abdullah Abdullah, Ameen, M. Sadeeq, & Zeebaree, 2021). Multimodal information fusion for integrating audio, visual, text and other information is crucial for MER. The conventional multimodal information fusion methods (Sidney K D'mello & Jacqueline Kory, 2015; Shoumy, Ang, Seng, Rahaman, & Zia, 2020) contain feature-level fusion, decision-level fusion, as well as model-level fusion, as described below.

Feature-level fusion, also known as early fusion (EF), is a relatively simple method which directly concatenates the extracted features from the single modality into a whole feature vector, and then feeds it into a classifier for emotion classification. However, such EF method is prone to suffer from "the curse of dimensionality" when concatenating multiple feature vectors. Moreover, EF does not take the temporal scales and thus fails to capture the associations across different modalities.

Decision-level fusion, also called as late fusion (LF), adopts a certain algebraic rule, such as "majority vote", "maximum", "sum", "minimum", "average", "product", *etc*., to merge the obtained results of various modalities into the final fusion results. The advantage of LF is that each modality can individually use a certain classifier that best suits the specific task for emotion classification. Nevertheless, LF considers different modalities to be independent of each other, thereby failing to reveal the relationships among different modalities.

Model-level fusion, has been widely used in recent years for emotion recognition tasks, which are designed to model each modality

individually while considering the correlation between modalities. As a result, it can take into account the interrelationships between different modes and reduce the need for these modal time synchronizations.

Hybrid-level fusion, is a combination of several different fusion strategies such as feature-level fusion, decision-level fusion as well as model-level fusion, thereby combining the advantages of different fusion strategies.

According to the amount of single-modal information employed, common MER methods can be divided into bimodal emotion recognition and trimodal emotion recognition. This section analyzes these fusion methods from the two aspects of emotion recognition based on bimodality and trimodality, and the summarizing results are shown in Table 6.

3.1. Bimodal emotion recognition

(1) Audio-visual Emotion Recognition

At present, there are several prevalent algorithms for audio-visual emotion recognition, such as CNN, DBN, GCN, Transformer, FCN, FBP, *etc*. The principles of the above-mentioned algorithms and technologies are as follows.

Zhang *et al.*, (Shiqing Zhang, Shiliang Zhang, Tiejun Huang, Wen Gao, & Qi Tian, 2018a) presented a hybrid deep learning approach for audio-visual emotion recognition. They initially extracted audio and video feature representations by using a CNN and a 3D-CNN, respectively. Next, a DBN model was used to merge the extracted audio and visual feature representations. Finally, emotion classification was performed by a linear SVM classifier (as shown in Fig. 3. (a)). Liu *et al.*, (Liu et al., 2021) developed a capsule graph convolutional network (CapsGCN) for audio-visual emotion recognition. Firstly, they extracted the spectrogram from speech signals, followed by 2D-CNNs for speech feature extraction (as shown in Fig. 3. (b)). They adopted a VGG16 to achieve visual features. Then, the extracted audio and visual feature representations were distilled by using capsule networks, and encapsulated into multimodal capsules, respectively. Next, graph convolutional networks (GCN) were utilized to learn the graph structure of multimodal capsules in terms of their inter-relations and intra-relations, and obtain hidden representations. Finally, these multimodal capsules as well as hidden relational representations achieved by CapsGCN were fed into a multi-head self-attention layer to enhance their discriminating power, followed by a FC layer and a Softmax layer for emotion classification. Fig. 4 shows the framework of audio-visual emotion recognition based on CapsGCN. Huang *et al.*, (Huang, Tao, Liu, Lian, & Niu, 2020) proposed to employ the Transformer model to integrate audio and visual modalities for emotion recognition (as shown in Fig. 3. (c)). For speech feature extraction, they obtained the typical acoustic feature set called eGeMAPS as speech features. For visual feature extraction, they extracted geometric features, such as facial landmark locations, facial action units, head postures, and so on. After encoding audio and visual modalities, a multi-head attention strategy was used to generate multimodal affective intermediate feature representations from the common semantic feature space, and then combined the Transformer with LSTMs for further performance promotion. They finally obtained continuous emotion recognition results through a linear FC layer. Zhou *et al.*, (Zhou et al., 2021) provided a multimodal attention fusion model on the basis of adaptive and multi-level factorized bilinear pooling (FBP) for audio-visual emotion recognition. For the audio modality, a fully convolutional network (FCN) with local response normalization was developed to learn high-level feature representations from speech spectrograms (as shown in Fig. 3. (d)). For the video modality, the pre-trained VGG-face model was utilized for extracting video feature representations. Then, a global FBP (G-FBP) strategy integrating self-attention was employed to conduct audio-visual information fusion. Moreover, an adaptive G-FBP (AG-FBP) strategy was presented to automatically compute the importance weights of audio and visual modalities when using G-FBP fusion. Finally, an adaptive and multi-level FBP (AM-FBP) approach was presented to combine global-trunk data as well as intra-trunk data on the top of AG-FBP. Based on the obtained fusion vector, a FC layer and a Softmax layer was adopted for emotion classification. Middya *et al.*, (Middya, Nag, & Roy, 2022) presented an optimal MER model by means of combining audio and visual modalities at different fusion strategies. Different CNN-based feature extractor networks, equipped with different convolutional layers as well as pooling layers, were designed for audio and visual feature extraction tasks (as shown in Fig. 3. (e)). Then, they concatenated the flattened audio-visual feature representations into a whole feature vector, followed by a FC layer and a Softmax layer for conducting final video emotion classification. They obtained better results at model-level fusion than feature-level and decision-level fusion. Sharafi *et al.*, (Sharafi, Yazdchi, Rasti, & Nasimi, 2022) developed a spatio-temporal CNN framework for audio-visual emotion recognition, in which audio and visual modalities were integrated as the input to a hybrid network consisting of a Bi-LSTM and two CNNs (as shown in Fig. 3. (f)). They concatenated the extracted spatio-temporal features from video frames, the extracted MFCCs and energy features from speech signals, and the Bi-LSTM network outputs into a long feature vector. Then, a FC layer and a Softmax layer was employed to identify emotion categories.

(2) Audio-text Emotion Recognition

Recently, there are several popular algorithms for audio-text emotion recognition (S. Zhang, Yang et al., 2023), such as CNN, RNN, LSTM, GRU, AIA-Net, and GCN, *etc*. The principles of the above-mentioned algorithms and technologies are as follows.

Hazarika *et al.*, (Hazarika, Gorantla, Poria, & Zimmermann, 2018) proposed a feature-level fusion approach integrating self-attention for audio-text emotion recognition (as shown in Fig. 4. (a)). High-dimensional hand-crafted LLDs like loudness, pitch, voice quality, MFCCs, *etc*., were extracted for speech signals. CNNs were leveraged to achieve textual features by means of extracting location-invariant local patterns. Then, they generated the fused vector by implementing a weighted addition based on these attention score values. Finally, a FC layer and a Softmax layer were utilized for emotion classification. The proposed feature-level fusion approach outperformed common fusion approaches like concatenation, outer-product, *etc*. Priyasad *et al.*, (Priyasad, Fernando, Denman, Sridharan, & Fookes, 2020) proposed to employ deep learning algorithms to fuse audio and text modalities for emotion recognition (as shown in Fig. 4. (b)). The SincNet layer (Ravanelli et al., 2018) was used to obtain speech features from original speech signals. Two sub-networks (a CNN, as well as a Bi-RNN followed by a CNN) were utilized for text feature extraction, in which the cross-attention mechanism was employed to learn the *N*-gram level correlation on hidden feature representations obtained by a Bi-RNN. Based on the concatenated audio and text features, the self-attention-based fusion method with a FC layer and a Softmax layer was used to obtain final emotion recognition results. Krishna *et al.*, (Krishna & Patil, 2020) developed a new MER approach based on cross-modal attention and 1D-CNNs (as shown in Fig. 4. (c)). They used an audio encoder (CNN + Bi-LSTM) to achieve high-level feature representations from the original speech signals, and a text encoder (Glove + CNN) to extract high-level text semantic information. Then, the cross-modal attention network was utilized to interactively fuse audio and text sequences, followed by a FC layer and a Softmax layer for emotion classification.

Likewise, Lian *et al.*, (Lian, Liu, & Tao, 2021) proposed to employ a conversational transformer model for conversational emotion classification (as shown in Fig. 4. (d)). The single-modal transformer as well as cross-modal transformer was presented to individually capture intra-modal and cross-modal interactions among various modalities. They adopted typical speech feature sets like eGeMAPS, and 300-dimensional word-level lexical features as their inputs for capturing temporal characteristics in the utterance. Besides, in order to model context-sensitive as well as speaker-sensitive dependencies, a Bi-GRU network and speaker embedding using the multi-head attention mechanism was used. Audio-text-speaker fusion component (ATS-Fusion) was provided for

Table 6
A summary for multimodal information fusion methods (ACC: Accuracy; CCC: Consistency Correlation Coefficient; WA: Weighted accuracy; UA: Unweighted accuracy; WAA: weighted average accuracy; UAA: unweighted average accuracy; WAF1: weighted average F1; UAF1: unweighted average F1; MAE: Mean Absolute Error; Corr: Correlation coefficient.).

Multimodal	Fusion	Publication	Features	Classifiers	Datasets	Performance
Audio-visual	Model-level	(Zhang et al., 2018a)	Audio: CNN Visual: 3D-CNN	Linear SVM	1)RML 2) eNTERFACE'05 3)BAUM-1 s	1)Acc(6-class):80.36% 2)Acc(6-class):85.97% 3)Acc(6-class):54.57%
	Model-level	(Huanget al., 2020)	Audio: LLDs Visual: Geometric Features	Linear FC	AVEC-2017	CCC(Arousal):0.654 CCC(Valence):0.708
	Model-level	(Liu et al., 2021)	Audio: CNN Visual: VGG16	FC + Softmax	eNTERFACE'05	1)Acc(6-class):80.83% 2)F1-score:80.23%
	Model-level	(Zhou et al., 2021)	Audio: FCN Visual: VGG-face	FC + Softmax	1)IEMOCAP 2)AFEW8.0	1)Acc(4-class):75.49% 2)Acc(7-class):63.09%
	Hybrid-level	(Middya et al., 2022)	Audio: CNN Visual: CNN	FC + Softmax	1)SAVEE 2)RAVDESS	1)Acc(7-class):99% 2)Acc(8-class):86%
	Feature-level	(Sharafi et al., 2022)	Audio: LLDs Visual:CNN + Bi-LSTM	FC + Softmax	1)SAVEE 2)RAVDESS 3)RML	1)Acc(7-class):99.75% 2)Acc(7-class):94.99% 3)Acc(7-class):99.23%
Audio-text	Feature-level	(Hazarika et al., 2018)	Audio: LLDs Text: CNN	FC + Softmax	IEMOCAP	Acc(4-class):71.4% ; F1-score:71.3%
	Model-level	(Priyasad et al., 2020)	Audio: SincNet + CNN Text: CNN + Bi-RNN	FC + Softmax	IEMOCAP	Acc(4-class):80.51% WA:79.22% UA:80.51%
	Model-level	(Krishna & Patil, 2020)	Audio: CNN + Bi-LSTM Text: Glove + CNN	FC + Softmax	IEMOCAP	Acc(4-class): 72.82%
	Model-level	(Lian et al., 2021)	Audio: LLDs Text: Lexical Features	FC + Softmax	1)IEMOCAP 2)MELD	1)IEMOCAP: WAA(4-class):83.6% WAF1(4class):83.8% WAA(6-class):68.0% WAF1(6class):67.5% 2)MELD: WAA(7-class):62.0% WAF1(7class):60.5%
	Model-level	(Zhang et al., 2022)	Audio:Wav-RoBERTa Text: RoBERTa	FC + Softmax	1)IEMOCAP 2)MELD 3)CMU-MOSEI	1)IEMOCAP: Acc(4-class):87.44% F1-score(4class): 87.16% 2)MELD: Acc(7-class):65.09% 3)CMUMOSEI:MAE:0.574 Acc(7-class):53.20% Acc(2-class):89.33% F1-score:89.33%
	Model-level	(Fu et al., 2022)	Audio: CNN + Bi-LSTM Text: BERT	FC + Softmax	1)IEMOCAP 2)MELD	1)IEMOCAP:Acc(4-class):85.82% F1-score:85.90% 2)MELD:Acc(7-class):66.40% F1-score:64.63%
Audio,visual, text	Model-level	(Poria, Cambria, Hazarika, Mazumder et al., 2017)	Audio: LLDs Visual: 3D-CNN Text: Word2vec	FC + Softmax	CMU-MOSI	Acc(2-class):81.3%
	Hybrid-level	(Pan et al., 2020)	Audio: LLDs Visual: 3D-CNN Text: Word2vec	FC + Softmax	IEMOCAP	Acc(4-class):73.94%
	Model-level	(Mittal et al., 2020)	Audio: LLDs Visual: Facial Landmarks Text: Glove	FC + Softmax	1)IEMOCAP 2)CMU-MOSEI	1)IEMOCAP: Acc(4-class):82.7 F1-score:82.40% 2)CMU-MOSEI:Acc(6-class):89.0% F1-score:90.2%
	Model-level	(Wang et al., 2020)	Audio: LLDs Visual:3D-CNN Text: CNN	FC + Softmax	1)IEMOCAP 2)CMU-MOSI 3)MELD	1)Acc(6-class):60.81% 2)Acc(2-class):82.71% 3)Acc(7-class):67.04%
	Decision-level	(Dai et al., 2021b)	Audio: CNN Visual: CNN Text: Transformer	FC + Softmax	1)IEMOCAP 2)CMU-MOSEI	1)IEMOCAP: Acc(6-class):84.4% F1-score:57.4% 2)CMU-MOSEI: Acc(6-class):66.8% F1-score:46.8%
	Model-level	(Ren et al., 2021)	Audio: LLDs Visual: 3D-CNN Text: Glove	FC + Softmax	IEMOCAP	Acc(6-class):65.0% F1-score: 64.5%
	Model-level	(Mai, Hu et al., 2022)	Audio:Bi-GRU Visual: Bi-GRU Text: Bi-GRU	FC + Softmax	1)IEMOCAP 2)CMU-MOSEI 3)CMU-MOSI 4)MELD	1)IEMOCAP:Acc(4-class):83.45% F1-score:82.63% 2)CMU-MOSEI: Acc(2-

(*continued on next page*)

Table 6 (*continued*)

Multimodal	Fusion	Publication	Features	Classifiers	Datasets	Performance
						class):82.4% Acc(7-class):50.9% F1-score:82.6% MAE:0.598 Corr:0.69 3)CMU-MOSI: Acc(2-class):82.3% Acc(7-class):39.4% F1-score:82.5% MAE:0.896 Corr:0.697
	Model-level	(Zheng et al., 2022)	Audio: AlexNet Visual: ResNet Text: Word2vec	FC + Softmax	1)IEMOCAP 2)MSP-IMPROV 3)CMU-MOSI 4)CMU-MOSEI	1)IEMOCAP: Acc(4-class):86.3 F1-score:86.5% 2)MSP-IMPROV: Acc(4-class):71.8% F1-score:71.8% 3)CMU-MOSI: Acc(2-class):85.2% Acc(7-class):46.6% F1-score:85.1% MAE:0.713 Corr:0.790 4)CMU-MOSEI:Acc(2-class):85.4% Acc(7-class):52.8% F1-score:85.6% MAE:0.601 Corr:0.776
	Model-level	(Mai, Zeng et al., 2022)	Audio: LLDs Visual: Facial Landmarks Text: Glove	Contrastive Learning	1)CMU-MOSI 2)CMU-MOSEI	1)Acc(7-class):48.3% 2)Acc(7-class):53.4%
	Model-level	(Zhao et al., 2022)	Audio: Wav2Vec2.0 Visual: DenseNet Text: BERT	FC + Softmax	1)IEMOCAP 2)MSP-IMPROV	1)IEMOCAP:WA(4-class):80.01% UAR(4-class):81.09% Acc(4-class):80.01% 2)MSP-IMPROV: WA(4-class):72.36% UAR(4-class):72.22% Acc(4-class):72.36%

audio-text fusion. Finally, a FC layer and a Softmax layer was utilized for emotion classification. Zhang *et al.*, (Zhang, Li, Chen, Yuan, & Chen, 2022) presented an adaptive interactive attention network (AIA-Net) for audio-text emotion recognition, in which a number of adaptive interactive attention collaborative learning layers were included (as shown in Fig. 4. (e)). In this AIA-Net, text cues were taken as a primary modality, whereas audio cues were considered as an auxiliary modality. The proposed AIA-Net initially adapted to text and speech feature representations at various dimensions, and captured their dynamic interactive interrelations adaptively. The dynamic interactive interrelations were encoded as different interactive attention weights, which were used to concentrate on the effective speech feature representations for textual emotional representations. For the text modality, they employed the pre-trained RoBERTa (Liu et al., 2019) to extract textual feature representations. For the audio modality, they adopted the pre-trained Wav-RoBERTa, equippted with 12 transformer layers and an output embedding layer, to extract speech features. Multiple collaborative learning layers were used for multiple multimodal interactions. Finally, after obtaining the shared audio-text features, a FC layer and a Softmax function was used for emotion classfication. Fu *et al.*, (Fu et al., 2022) developed an audio-text emotion recognition method based on context and knowledge-aware graph convolutional networks (GCN) (as shown in Fig. 4. (f)). For the audio modality, they used a CNN to obtain deep speech features from segment-level spectrograms, followed by a Bi-LSTM for temporal modelling in an utterance. For the text modality, they fine-tuned the pre-trained BERT model on target datasets to achieve textual features from the transcripts. Based on the extracted auido and text features, they initially created two different graphs to model the contextual interaction and knowledge dynamical information. Next, they embedded an emotional lexicon into the building knowledge graphs. Then, they obtained a balance between the contexts as well as emotion-enriched knowledges, which were incorporated into a newly-developed adjacency matrix in a GCN. Finally, they jointly taught them with audio and text modalities in order to effectively capture the semantics-sensitive as well as knowledge-sensitive contextual dependencies between utterances in each conversation. Subsequently, a FC layer and a Softmax function was leveraged for emotion recognition.

(3) Visual-text Emotion Recognition

Recently, there are several popular algorithms for visual-text emotion recognition, such as CNN, R-CNN, and hybrid models, *etc*. The principles of the above-mentioned algorithms and techniques are as follows.

Poria *et al.*,(Poria, Chaturvedi, Cambria, & Hussain, 2016) developed a temporal CNN to extract both visual and textual features from different modalities. The CNN model combines each temporal pair of images to a single image, which makes the model sensitive to sequence. The model learns a dictionary of features across languages. Kumar *et al.*, (Kumar & Garg, 2019) proposed a multimodal emotion recognition model extracting textual and image features to analyze sentiment polarity. The image sentiment analysis and text sentiment analysis were implemented through R-CNN and context-aware hybrid model, respectively. Later, Kumar *et al.*, (Kumar, Srinivasan, Cheng, & Zomaya, 2020) used textual and visual modalities of social data in the emotional analysis process, as well as text embedded along with an image to extract features. They proposed a deep learning classification model called Hybrid Context ConvNet SVMBoVW framework, which includes text analysis module,

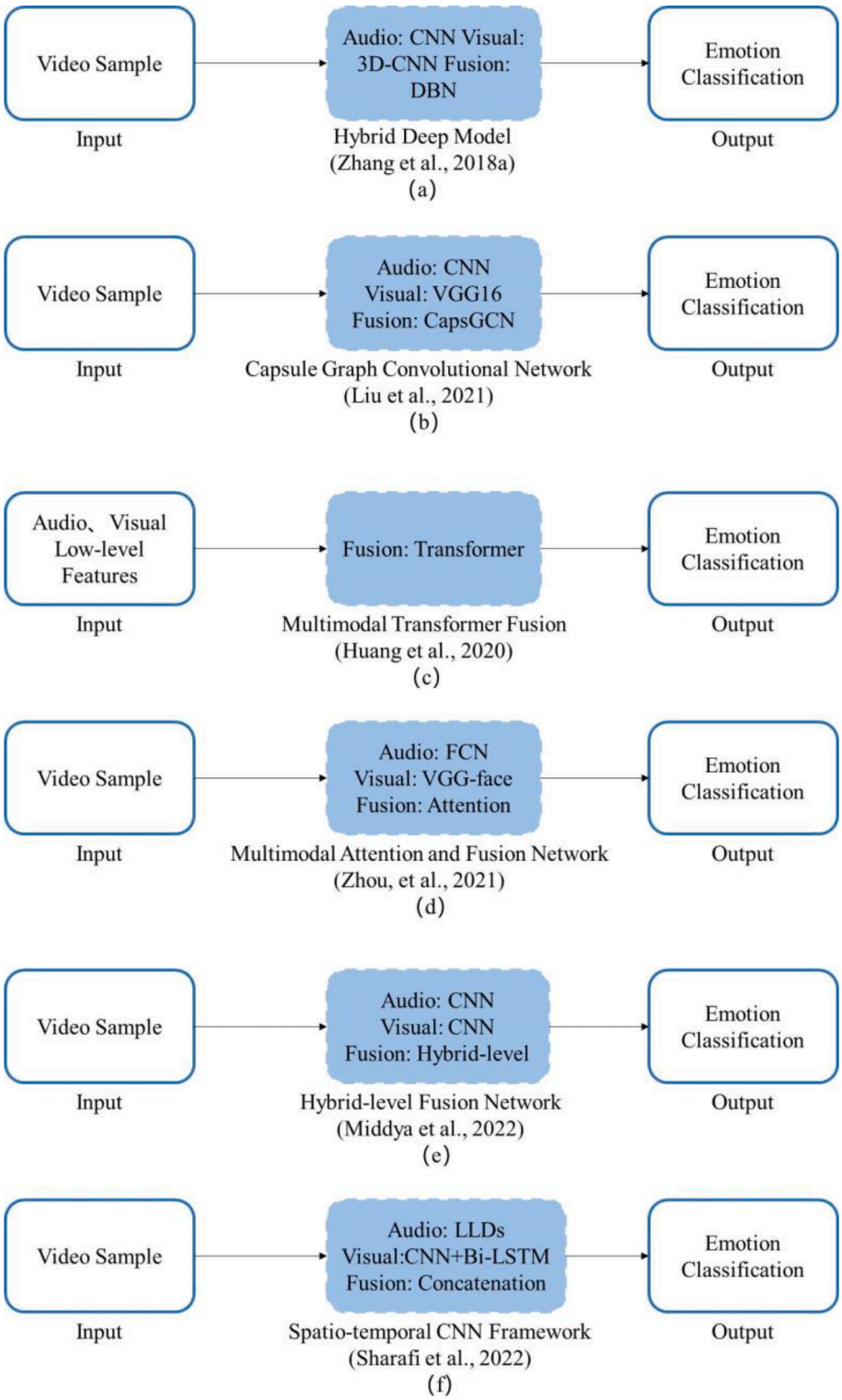


Fig. 3. A summary of the framework of various audio-visual emotion recognition including audio-visual feature extraction and fusion methods.

image analysis module, and other modules. This design can enhance the emotional analysis of different modalities in online social media. Xu *et al.*, (B. Xu, Fu, Jiang, Li, & Sigal, 2018) proposed a framework to solve the heterogeneous external sources transferring problem, the framework can learn video encoding from auxiliary emotion image dataset, and then transfer knowledge from auxiliary textual data to provide information for emotion recognition which are unseen during training.

Highlights in this section: (1)In the process of audio-visual and audio-text fusion for MER, model-level fusion has been the most popular one due to its ability of capturing the inter-correlation across different modalities. (2) For audio modality, typical LLD features such as eGeMAPS, or CNN-based learned features have become the predominant speech features for emotion classification. (3) For visual modality, CNN-based pre-trained models like VGG-face and VGG16, or CNN + Bi-LSTM, have been widely utilized for deep visual feature extraction. (4) For text modality, pre-trained word embeddings such as Word2vec, Glove and

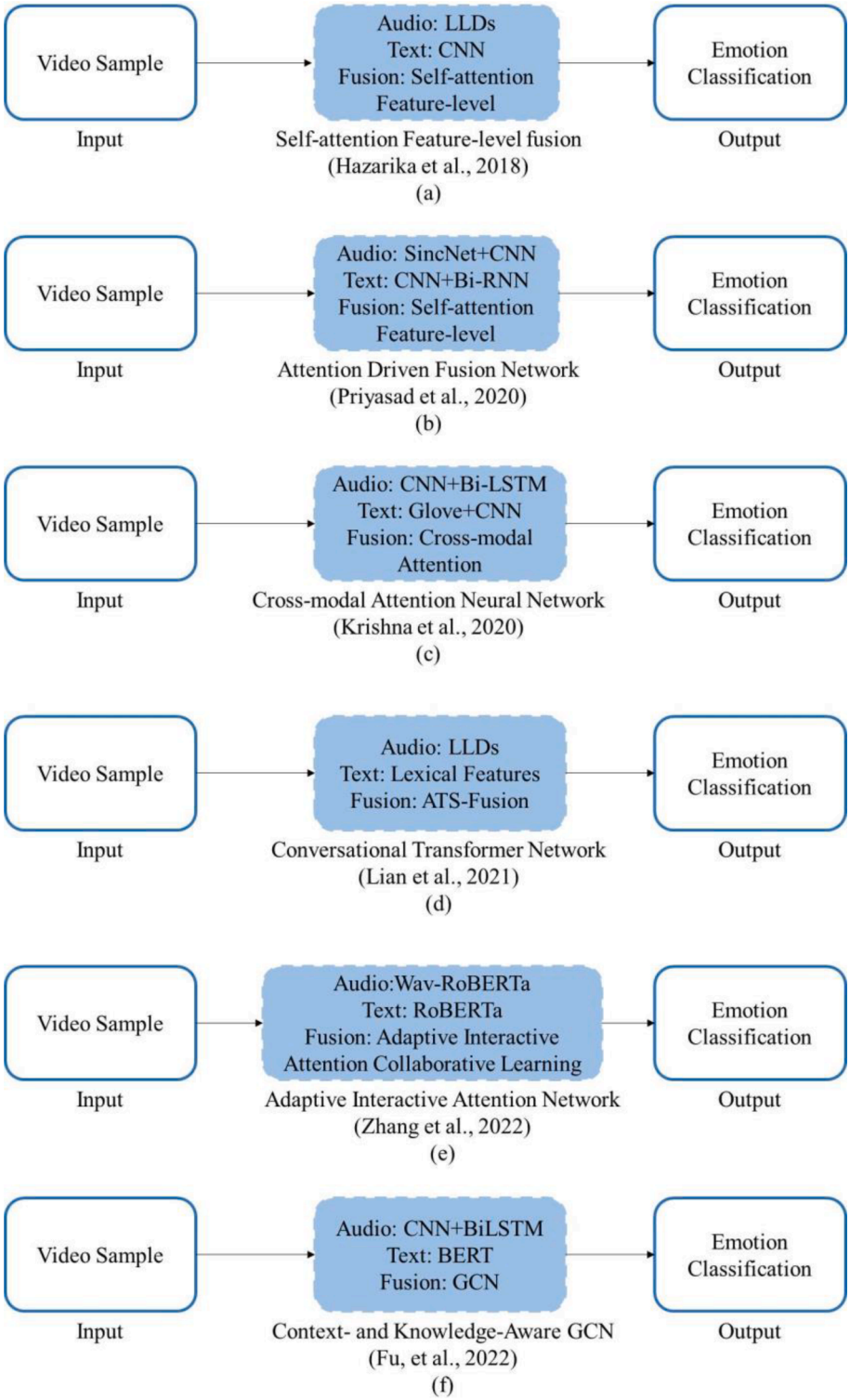


Fig. 4. A summary of the framework of various audio-text emotion recognition including audio-text feature extraction and fusion methods.

BERT, or CNN + Bi-RNN, have been usually adopted to extract text semantic representations.

3.2. Trimodal emotion recognition

More recently, a large number of researchers have begun to pay attention to the trimodal emotion recognition problem and proposed a variety of algorithms, i.e., MESM, CAT-LSTM, MMAN, M3ER, Bi-GRU, TransModality, IMAN, and MEmoBERT, *etc*. The principles of the above-mentioned algorithms are as follows.

Dai *et al.*, (Dai et al., 2021b) presented a multimodal end-to-end sparse model (MESM) to perform emotion classification. For audio and visual modalities, the pre-trained VGG16 model was used for initial audio and visual feature extraction tasks, followed by a Transformer to model the temporal information (as shown in Fig. 5. (a)). For text modality, a Transformer was directly employed for encoding the sequence of words. To reduce the computational cost, a cross-modal sparse CNN block, consisting of a sparse CNN and a cross-modal attention block, was

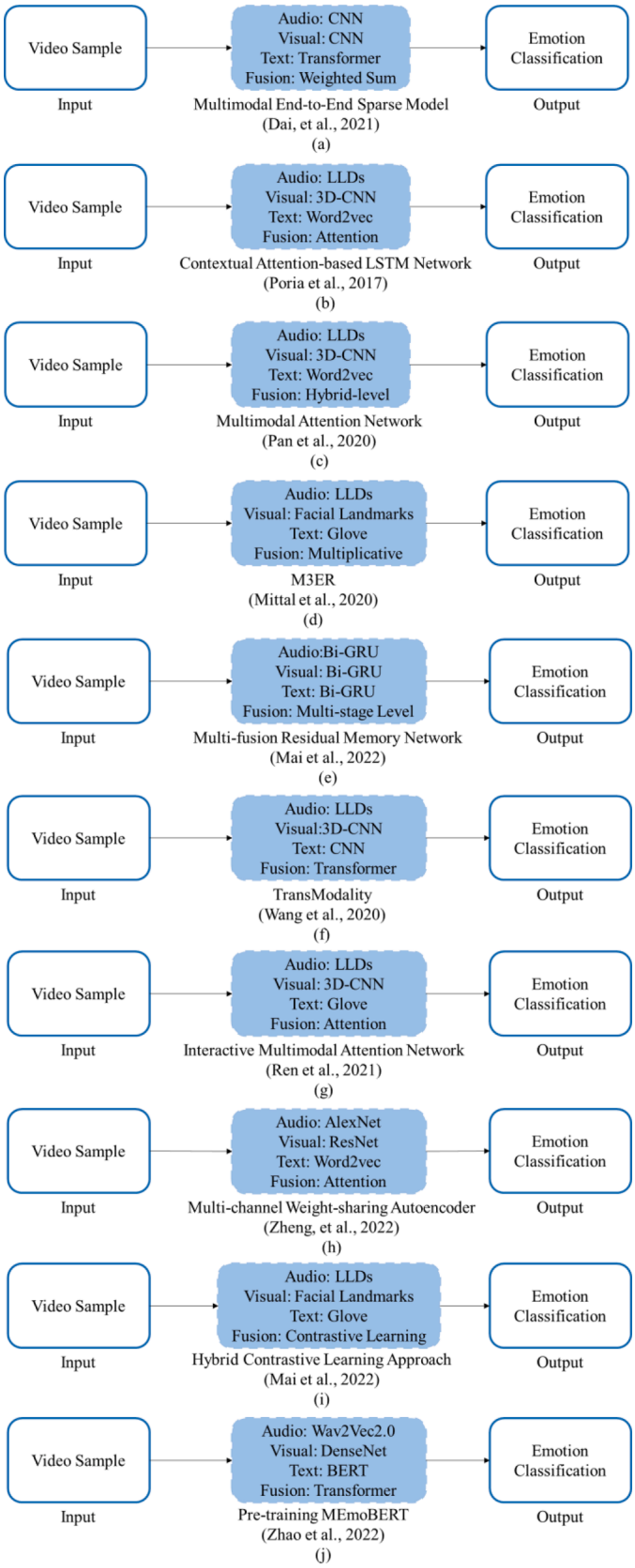


Fig. 5. A summary of the framework of various trimodal emotion recognition including trimodal feature extraction and fusion methods.

introduced for audio and visual feature extraction. Finally, the classification results were obtained by feed-forward network (FFN) containing a linear FC layer and a Softmax layer. A weighted sum of the obtained classification scores from audio, visual, text modalities was performed to produce final emotion prediction scores. Poria *et al.*, (Poria, Cambria, Hazarika, Mazumder et al., 2017) provided a contextual attention-based LSTM (CAT-LSTM) network to capture contextual information in utterances for contextual multimodal sentiment analysis. For audio modality, they extracted LLD features such as sound intensity, pitch and their statistics (as shown in Fig. 5. (b)). For visual modality, they employed a 3D-CNN to capture video features. For text modality, they utilized the pre-trained Word2vec to obtain textual features. An attention-based fusion mechanism was adopted for multimodal classification fusion, followed by a FC layer and a Softmax layer for final emotion classification. Pan *et al.*, (Pan, Luo, Yang, & Li, 2020) proposed a hybrid fusion strategy called multi-modal attention network (MMAN) for MER. They employed typical LLD features (ComParE-2013) as speech features, a 3D-CNN for extracting visual features, a Word2vec for extracting textual features (as shown in Fig. 5. (c)). For each modality, a contextual long short-term memory (cLSTM) model with a LSTM layer and two FC layers was used to learn the contextual information in utterances from a conversation. The presented MMAN was comprised of one multi-modal attention cLSTM for early fusion, three independent unimodal blocks such as cLSTM-text, cLSTM-visual as well as cLSTM-speech. The classification results of these four sub-networks were integrated with a FC layer and a Softmax layer for late fusion. Mittal *et al.*,(Mittal et al., 2020) developed a multimodal emotion recognition model called M3ER using a multiplicative fusion layer, which aimed to learn more reliable modalities and suppress others on a sample basis (as shown in Fig. 5. (d)). They employed popular LLD features like MFCCs, pitch, glottal source parameters as speech features, and common facial action units and facial landmarks as visual features, respectively. For text modality, the pre-trained Glove was used to derive texture features. Then, they performed a multiplicative modality fusion scheme, followed by a FC layer and a Softmax layer for emotion classification.

Likewise, Mai *et al.*, (Mai, Hu, Xu, & Xing, 2022) proposed a multi-fusion residual memory network approach for identifying utterance-level emotions (as shown in Fig. 5. (e)). For each modality, they adopted a bidirectional GRU (Bi-GRU) model to learn temporal interaction cues among time steps as well as achieve context-dependent representations. Then, several FC layers were used to further learn intra-modality interaction cues. These extracted features for each modality were fed into a multi-stage fusion module consisting of an attention-guided fusion stage as well as a time-step level fusion stage. Finally, the emotion classification results were obtained by a FC layer and a Softmax layer. Wang *et al.*, (Z. Wang, Wan, & Wan, 2020) presented an end2end fusion approach with Transformer named TransModality for MER (as shown in Fig. 5. (f)). They leveraged a CNN, a 3D-CNN and the conventional OpenSMILE tool to extract typical textual, visual and audio features, respectively. With TransModality, the learned features reflected the information of source and target modalities. Then, they fed the learned features into two modality fusion cells for text-visual and text-audio modality fusion, respectively. Finally, a FC layer and a Softmax layer was used for final emotion classification. Ren *et al.*, (Ren, Huang, Shi, & Nie, 2021) designed an interactive multimodal attention network (IMAN) approach to perform MER tasks in conversation (as shown in Fig. 5. (g)). This IMAN extracted typical speech features called ComParE-2013 for the audio modality. They employed a 3D-CNN to derive spatio-temporal features across frames for the visual modality, and a simple CNN block including a convolutional layer and a max-pooling layer to obtain text features for text modality, respectively. In this IMAN, two modules such as a cross-modal attention fusion module as well as a conversational modeling module were included. The former module aimed at capturing cross-modal interaction cues of multimodal information. The latter module focused on learning the contextual cues and speaker dependencies in conversation. Finally, the classification

results were obtained through a FC layer and a Softmax layer. Zheng *et al.*, (Zheng, Zhang, Wang, Wang, & Zeng, 2022) proposed a new MER approach of multi-channel weight-sharing Autoencoder integrating cascade multi-head attention (as shown in Fig. 5. (h)). Specifically, they extracted RGB-like image Mel-spectrograms as inputs of traditional CNNs like AlexNet to derive speech features. They employed the pre-trained ResNet to extract visual features, and the pre-trained Word2-vec to achieve text features. Next, a scalable heterogeneous feature fusion block was constructed by combining several different multi-head attention modules in series. Finally, based on the concatenation as well as convolutional operations, all extracted features were merged into a whole feature vector, followed by a FC layer and a Softmax layer for emotion identification.

In addition, Mai *et al.*, (Mai, Zeng, Zheng, & Hu, 2022) presented a hybrid contrastive learning approach of trimodal feature representations for multimodal sentiment analysis (as shown in Fig. 5. (i)). For the audio modality, COVAREP (Degottex, Kane, Drugman, Raitio, & Scherer, 2014) was used to extract typical LLDs such as MFCCs, pitch tracking, glottal closure instants, spectral envelope, and so on. For the visual modality, they extracted a set of hand-crafted features such as facial landmarks, head poses, and so on. For the text modality, the pre-trained Glove was employed to achieve high-level text feature representations. Then, they performed intra-/inter-modal contrastive learning as well as semi-contrastive learning so as to capture inter-sample and inter-class inter-correlation. Zhao *et al.*, (Zhao, Li, Jin, Wang, & Li, 2022) developed a pre-training deep model called MEmoBERT for MER. MEmoBERT was capable of learning multimodal joint feature representations by means of a self-supervised learning scheme from large-scale unlabeled video data (as shown in Fig. 5. (j)). MEmoBERT was composed of three modality-specific encoder blocks, three modality-specific embedder blocks as well as a multi-layer cross-modal Transformer block. For the audio modality, the pre-trained Wav2Vec2.0 (Baevski, Zhou, Mohamed, & Auli, 2020) was utilized to obtain frame-level acoustic features. For the visual modality, the pre-trained DenseNet was leveraged to obtain facial expression-related features. For the text modality, the pre-trained BERT was utilized to obtain text features. Then, a multi-layer Transformer block was employed to capture cross-modality contextual feature representations across various modalities, followed by a FC layer and a Softmax layer for emotion classification.

Highlights in this section: (1) As shown in Table 6, when fusing audio, visual and text modalities, model-level fusion has become the mainstream method, since it aims to model the inter-correlation across different modalities. In particular, the cross-modal attention-based fusion mechanism has been a promising direction for multimodal information fusion in recent years. (2) For feature extraction, the pre-trained deep learning models have been popular. More specially, for the audio modality, CNN-based pre-trained models, such as AlexNet, Wav-RoBERTa, and Wav2Vec2.0, have exhibited excellent performance for audio feature extraction. For the visual modality, CNN + Bi-LSTM or several pre-trained models, like ResNet, VGG16, VGG-face, DenseNet, and 3D-CNN, have been predominant for visual feature extraction tasks. For the text modality, the pre-trained CNN-based word embeddings, like Word2vec, Glove, BERT, and RoBERTa, have been the popular methods for text feature extraction.

4. Research challenges and open issues

4.1. Lightweight and explainable deep models for MER

Recently, various deep learning approaches have been successfully employed for learning high-level feature representations for MER. Furthermore, these deep learning approaches usually outperform hand-crafted features. However, most of existing deep learning models have massive network parameters, leading to their high computational complexity. In this case, it is definitely difficult to employ such complicated deep learning models for real-time HCI applications, especially in mobile phone devices. In order to alleviate this problem, tremendous works concentrate on model compression and acceleration of deep neural networks in an effort to produce lightweight deep learning models for real-time applications. To this end, there are four groups of representative approaches (Cheng, Wang, Zhou, & Zhang, 2018) to conduct model compression and acceleration: parameter pruning and sharing, low-rank factorization, transferred/compact convolutional filters, as well as knowledge distillation.

For instance, Zheng *et al.*, (Zheng, Chen, Ding, & Luo, 2022) developed a differentiable network channel pruning approach for model compression. Specifically, they pruned the network in terms of the learnable probability to achieve the optimal substructure, once the network parameters were optimized. Lin *et al.*,(Lin, Ji, Chen, Tao, & Luo, 2019) presented a holistic CNN compression approach, in which a low-rank decomposition strategy was developed to reduce redundancies across convolutional kernels as well as fully-connected matrices.

Although deep learning methods have yielded promising performance on different feature learning tasks, the black-box problem still exists. More specially, the training procedure of deep learning models usually lacks transparency, leading to the fact that it is hard to understand what kind of internal representations have been learned by deep learning models during the training procedure. To mitigate this issue, a variety of interpretable methods on deep learning models have been presented in recent years. In general, existing interpretable methods can be grouped into two categories: global and local interpretation methods (Liang, Li, Yan, Li, & Jiang, 2021). Global interpretation approaches aim to provide an overall understanding of the learned components like the network weights and structures, whereas local interpretation approaches aim to check individual predictions locally. In this case, local interpretation approaches are much easier for implementation and have lower computation complexity than global interpretation approaches. According to the implementation manner of local interpretation approaches, local interpretation approaches can be grouped into: data-driven approaches (Zia, Bashir, Ullah, & Murtaza, 2022) relying on input data for interpretation, and model-driven approaches (Belharbi et al., 2022) analyzing the internal components (such as the weights) of a deep learning model. Especially, feature visualization based on gradient-weighted class activation mapping (Grad-CAM) (Selvaraju et al., 2017) has been one of the predominant model-driven approaches to show the learned emotional features of deep learning models for MER (D. Li et al., 2022).

4.2. Multimodal information fusion strategies for MER

How to effectively fuse different modalities is a crucial step for MER. The representative multimodal information fusion strategies (Sidney K D'mello & Jacqueline Kory, 2015; Shoumy et al., 2020) are comprised of feature-level fusion, decision-level fusion, as well as model-level fusion. As shown in Table 6, it is found that model-level fusion has been the most popular method for MER. This is attributed to the advantages of model-level fusion methods compared with other two fusion methods. In particular, recently-developed cross-modal attention-based fusion strategies (Poria, Cambria, Hazarika, Majumder et al., 2017; Zhang et al., 2022; Jiahao Zheng, Zhang, Wang, & Zeng, 2022) have become a promising direction, since they are capable of effectively capturing the inter-correlation across different modalities.

Feature-level fusion like concatenation is the simplest for implementation among the above-mentioned fusion methods. Nevertheless, feature-level fusion is unable to capture the associations across various modalities, and thus its performance is heavily affected by time scales as well as metric levels of extracted features from various modalities. By contrast, decision-level fusion aims at modeling independently each input modality, and then integrating these results of single-modal emotion recognition with algebraic rules. However, decision-level fusion fails to capture the relationships among features of different modalities. In consequence, among three basic fusion methods, model-

level fusion methods usually perform best, thereby deserving a depth-in study on developing more advanced model-level fusion strategies for MER.

In addition, most of existing MER works in Table 6 fails to take advantage of the useful temporal and semantic alignment information across different modalities for MER. More specially, given a sequence of audio frames in an audio-text emotion recognition system, in which a speaker expresses "joy" in the utterance "That's great!", it is beneficial to concentrate on the word "great" and its corresponding audio frames in purpose of learning a more discriminating feature representation for MER. In this case, it is needed to perform a multimodal alignment for exploring the inherent temporal and semantic consistency to reconcile emotional-related information across modalities. Recently, a few works (Chen et al., 2021; Hou, Zhang, & Lu, 2022) have been tried to alleviate this problem in an audio and text emotion recognition system. Therefore, it is an interesting subject for other MER systems based on audio-visual, or audio-visual-text modalities to further investigate the effective mechanism to capture temporal and semantic alignment information across different modalities.

4.3. Cross-corpus MER

Although the above-mentioned MER studies have made great progress in a single-corpus setting, in which the training corpus and testing corpus originates from the same corpus, it remains a challenging issue in a cross-corpus setting. Nevertheless, existing MER models which are trained and tested in a single-corpus setting could not work well when dealing with new corpora due to the corpus-bias problem. In other words, existing MER models suffer from dramatic performance degradation, since feature distributions among different corpora have a big discrepancy owing to the differences in recording device quality, spoken languages, culture, emotion annotation, *etc*.

To address the above-mentioned issue, at present tremendous efforts have been devoted to conducting cross-corpus single-modal emotion recognition such as cross-corpus audio and facial expression emotion recognition (Chen, Song, & Zheng, 2021; Zhang, Liu, Tao, & Zhao, 2021). However, few studies involve in cross-corpus MER. Only one previous paper (Liang et al., 2019) tried to mitigate the cross-culture discrepancy for MER. In (Liang et al., 2019), they presented an adversarial learning framework, in which emotion recognition and culture recognition was considered as two adversarial tasks, so as to alleviate the culture influence on MER tasks. In future, it is an important direction to further explore GANs (Saxena & Cao, 2021) based adversarial learning methods for cross-corpus MER.

4.4. More modalities for MER

Previous MER works focus on assessing human affective states relying on audio, visual (facial expression) as well as text modalities. However, these types of input data lack sufficient objective features to accurately characterize a person's emotional states. Therefore, integrating more modalities is an interesting direction for MER.

In the past two decades, physiological signals based emotion recognition (Shu et al., 2018) has drawn much interest, since physiological signals are usually more reliable and objective for continuous real-time monitoring of a person's emotional states. The representative physiological signals contain electroencephalogram (EEG), electrocardiogram (ECG), galvanic skin response (GSR), and so on. Among these types of physiological signals, EEG signals (Alarcão & Fonseca, 2019; García-Martínez, Martinez-Rodrigo, Alcaraz, & Fernández-Caballero, 2019) have been the predominant one for emotion recognition. However, since EEG signals are nonlinear and non-stationary, the traditional linear methods based on statistical and frequency features (Taran & Bajaj, 2019) may not be suited well. In this sense, it is a meaningful direction to explore advanced nonlinear methods (García-Martínez et al., 2019; Rahman, Sarkar, Hossain, & Moni, 2022) of processing EEG signals for emotion recognition. Besides, each type of physiological signal has its advantages and disadvantages. Accordingly, it is very interesting to investigate how to effectively integrate different types of physiological signals (Zhang et al., 2021) for implementing multimodal (Tan, Sun, Duan, Solé-Casals, & Caiafa, 2021; Wang, Wang, Yang, & Zhang, 2022; Yang et al., 2021) physiological signals based emotion recognition. In addition, combining physiological signals with other modalities is also an important direction for promoting MER performance.

Except for facial expression, body language such as body gestures, postures, eye movement and so on, is another visual modality characterizing human emotion expression. Compared with facial expression, body language is much easier to suffer from the influence of gender differences and culture dependence (Noroozi et al., 2021). Additionally, there is a scarcity of body language-related emotional datasets with labeled categories. These two factors make body language-based emotion recognition remain a less explored topic. In recent years, constructing multimodal emotional datasets integrating body language with other modalities (J. Chen et al., 2021; Sapiński et al., 2019), such as facial expression, audio, and so on, has attracted extensive attentions. In this sense, based on these constructed multimodal emotional datasets, exploring the performance of MER integrating body language with other modalities deserves a depth-in study in future.

4.5. Few-shot learning for MER

Existing MER studies usually rely on a large number of annotated multimodal emotional samples, especially when using deep learning techniques for feature extraction. However, it is really hard to collect massive annotated multimodal emotional samples due to the expensive cost of manpower and resources. In addition, recent advances in psychology (Demszky et al., 2020) have provided new conceptual and computational methods to represent the complex "semantic space" of emotion with a fine-grained taxonomy, in which 27 distinct emotional categories are classified. In this case, emotion categories have been increasingly diverse and fine-grained, making it more difficult to collect annotated emotional samples.

To alleviate the above-mentioned issue, it is desirous to explore effective few-shot learning strategies for MER. A few recent works (Yu & Zhang, 2022; Y. Yu, Zhang, & Li, 2022) have tried to investigate the few-shot learning scenario for MER. More specially, Yu *et al.*, (Yu et al., 2022) proposed a prompt-based multimodal fine-tuning method to perform few-shot MER tasks. They proposed a unified pre-training strategy with two stages for the proposed method so as to bridge the semantic gap between text and visual modalities. In this sense, developing suitable multimodal pre-training strategies for few-shot MER is a very interesting subject in future. Similarly, investigating advanced zero-shot learning approaches (Qi, Yang, & Xu, 2021) for MER, which aims to identify rare unseen emotions, is also an important direction.

To sum up, the above-mentioned open issues in MER and possible solutions are listed in Table 7.

5. Summary & conclusions

In this work, we aim to systematically analyze and summarize the state-of-the-art DL-MER methods, including multimodal emotional datasets, hand-crafted and deeply-learned feature extraction methods related to audio, visual, and text modalities, multimodal information fusion strategies such as feature-level, decision-level as well as model-level fusion. In addition, several challenges and open issues are highlighted for further exploration.

Although existing emotion recognition works in unimodal or multimodal manners have made significant breakthroughs, the developed robust and effective approaches in diverse and challenging scenario are still very limited. Accordingly, we can conclude this review with several vital recommendations for future prospects in MER:

Table 7
A summary of open issues and possible solutions in MER.

Open issues	Possible solutions
Lightweight and Explainable Deep Models for MER	Lightweight: parameter pruning and sharing, low-rank factorization, transferred/ compact convolutional filters, as well as knowledge distillation (Cheng et al., 2018) Explainable: global and local interpretation methods (Liang et al., 2021).
Multimodal Information Fusion Strategies for MER	Feature-level fusion: concatenation Decision-level fusion: modeling independently each input modality, then integrating these results of single-modal emotion recognition Model-level fusion: cross-modal attention-based fusion strategies (Poria, Cambria, Hazarika, Majumder et al., 2017; Zhang et al., 2022; Jiahao Zheng, Zhang, Wang, & Zeng, 2022)
Cross-corpus MER	Cross-corpus audio and facial expression emotion recognition (Chen, Song, & Zheng, 2021; Shiqing Zhang, Liu, Tao, & Zhao, 2021 Mitigate the cross-culture discrepancy for MER by adversarial learning framework (Liang et al., 2019) GANs based adversarial learning methods (Saxena & Cao, 2021)
More Modalities for MER	Physiological signals (Noroozi et al., 2021) Body language such as body gestures, postures, eye movement and so on (Noroozi et al., 2021)
Few-shot Learning for MER	Prompt-based multimodal fine-tuning method (Yu et al., 2022) Advanced zero-shot learning approaches (Qi et al., 2021)

(1) Due to the scarcity of large-scale multimodal emotion datasets, it is urgently needed to construct large-scale annotated multimodal emotional datasets so as to provide a strong support for the training of deep learning models. Additionally, developing new multimodal emotional datasets, which integrates common audio, visual, and text modalities with other modalities, like physiological signals, body language, and so on, is an important direction. Moreover, based on the constructed multimodal emotional datasets with more modalities, it is possible and meaningful to carry out integrating more modalities for MER.

(2) Considering the high computational complexity of most deep leaning methods, it is desirous to explore promising deep model compression and acceleration techniques for developing light-weight deep learning models, which is very important in real-time HCI applications. Moreover, to address the black-box problem of deep learning methods, developing advanced explainable deep learning models, which aims to understand what kind of internal representations is truly learned by deep learning models during the training procedure, deserves a depth-in study.

(3) Although deeply-learned features have frequently shown better performance than hand-crafted features on MER tasks, the advantages of hand-crafted features cannot be neglected for MER tasks. This is because both of them may have their own advantages and disadvantages. Therefore, how to effectively learn the complementarity between hand-crafted as well as deep-learned features for further improving performance, is a research direction.

(4) When implementing multimodal information fusion for MER, it is important to capture the inter-correlation across various modalities. To this end, developing cutting-edge model-level fusion strategies is crucial for MER. In particular, current cross-modal attention-based fusion methods worth further improvement. Besides, few works take into account the useful temporal and semantic alignment information across different modalities for MER. Accordingly, how to effectively capture temporal and semantic alignment information across different modalities for MER, is an interesting issue.

(5) To alleviate the problem of insufficient multimodal emotional data, it is of great importance to develop effective few-shot learning strategies for MER. More specially, recently-emerged multimodal pre-training strategies for few-shot MER have been an important guiding direction. In addition, few works involve in cross-corpus MER. Therefore, it is also interesting to explore advanced cross-corpus learning methods for cross-corpus MER tasks.

(6) The problem of modality-missing is another relatively new open issue in MER tasks, since modality-missing usually emerges in real-world scenery. To address this problem, there are several techniques for providing potential solutions such as Bayesian meta-learning (Ma et al., 2021), maximum likelihood estimation (Ma, Xu, Huang, & Zhang, 2021), and a unified model with missing modality imagination network (Jinming Zhao, Li, & Jin, 2021).

CRediT authorship contribution statement

Shiqing Zhang: Conceptualization, Methodology, Resources, Writing – original draft, Supervision, Funding acquisition. **Yijiao Yang:** Validation. **Chen Chen:** Software, Investigation. **Xingnan Zhang:** Data curation. **Qingming Leng:** Visualization. **Xiaoming Zhao:** Formal analysis, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

This work was supported by National Natural Science Foundation of China (NSFC), and Zhejiang Provincial National Science Foundation of China under Grant No. 62276180, 61976149, 62066021, 62206117 and LZ20F020002.

References

Abbaschian, B. J., Sierra-Sosa, D., & Elmaghraby, A. (2021). Deep learning techniques for speech emotion recognition, from databases to models. *Sensors, 21*, 1249.

Abdul-Mageed, M., & Ungar, L. (2017). Emonet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 718-728). Vancouver, Canada.

Abdullah, S. M. S. A., Ameen, S. Y. A., Sadeeq, M., & Zeebaree, S. (2021). Multimodal emotion recognition using deep learning. *Journal of Applied Science and Technology Trends, 2*, 52–58.

Abdullah, S. M. S. A., Ameen, S. Y. A., Sadeeq, M. A., & Zeebaree, S. (2021). Multimodal emotion recognition using deep learning. *Journal of Applied Science and Technology Trends, 2*, 52–58.

Acheampong, F. A., Nunoo-Mensah, H., & Chen, W. (2021). Transformer models for text-based emotion detection: A review of BERT-based approaches. *Artificial Intelligence Review, 54*, 5789–5829.

Ahmed, M. R., Islam, S., Islam, A. M., & Shatabda, S. J. E. S. w. A. (2023). An ensemble 1D-CNN-LSTM-GRU model with data augmentation for speech emotion recognition. *218*, 119633.

Ahsan, T., Jabid, T., & Chong, U.-P. (2013). Facial expression recognition using local transitional pattern on Gabor filtered facial images. *IETE Technical Review, 30*, 47–52.

Akçay, M. B., & Oğuz, K. (2020). Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication, 116*, 56–76.

Akhtar, M. S., Ekbal, A., & Cambria, E. (2020). How intense are you? Predicting intensities of emotions and sentiments using stacked ensemble [application notes]. *IEEE Computational Intelligence Magazine, 15*, 64–75.

Alarcão, S. M., & Fonseca, M. J. (2019). Emotions recognition using EEG signals: a survey. *IEEE Transactions on Affective Computing, 10*, 374–393.

Amiriparian, S., Gerczuk, M., Ottl, S., Cummins, N., Freitag, M., Pugachevskiy, S., Baird, A., & Schuller, B. (2017). Snore Sound Classification Using Image-Based Deep Spectrum Features. In *Proc. Interspeech 2017* (pp. 3512-3516). Stockholm, Sweden.

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems, 33*, 12449–12460.

Bahreini, K., Nadolski, R., & Westera, W. (2016). Towards multimodal emotion recognition in e-learning environments. *Interactive Learning Environments, 24*, 590–605.

Baltrušaitis, T., Mahmoud, M., & Robinson, P. (2015). Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)* (Vol. 6, pp. 1-6). Ljubljana, Slovenia: IEEE.

Bänziger, T., Grandjean, D., & Scherer, K. R. (2009). Emotion recognition from expressions in face, voice, and body: The Multimodal Emotion Recognition Test (MERT). *Emotion, 9*, 691.

Bao, S., Xu, S., Zhang, L., Yan, R., Su, Z., Han, D., & Yu, Y. (2012). Mining social emotions from affective text. *IEEE Transactions on Knowledge and Data Engineering, 24*, 1658–1670.

Belharbi, S., Sarraf, A., Pedersoli, M., Ben Ayed, I., McCaffrey, L., & Granger, E. (2022). F-cam: Full resolution class activation maps via guided parametric upscaling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 3490-3499).

Bengio, Y., Lamblin, P., Popovici, D., & Larochelle, H. (2007). Greedy layer-wise training of deep networks. In *Advances in neural information processing systems* (pp. 153-160). Vancouver, B.C., Canada: MIT Press.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Bosch, A., Zisserman, A., & Munoz, X. (2007). Representing shape with a spatial pyramid kernel. In *Proceedings of the 6th ACM international conference on Image and video retrieval* (pp. 401-408). New York,NY,United States: ACM.

Bottou, L. (2012). *Stochastic gradient descent tricks. In Neural networks: Tricks of the trade* (pp. 421–436). Springer.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., … Askell, A. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., & Weiss, B. (2005). A database of German emotional speech. *Interspeech, 5*, 1517–1520.

Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., … Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation, 42*, 335–359.

Cai, R., Qin, B., Chen, Y., Zhang, L., Yang, R., Chen, S., & Wang, W. (2020). Sentiment analysis about investors and consumers in energy market based on BERT-BiLSTM. *IEEE Access, 8*, 171408–171415.

Canal, F. Z., Müller, T. R., Matias, J. C., Scotton, G. G., de Sa Junior, A. R., Pozzebon, E., & Sobieranski, A. C. (2022). A survey on facial emotion recognition techniques: A state-of-the-art literature review. *Information Sciences, 582*, 593–617.

Chen, D., Song, P., & Zheng, W. (2021). Learning transferable sparse representations for cross-corpus facial expression recognition. *IEEE Transactions on Affective Computing*. https://doi.org/10.1109/TAFFC.2021.3077489

Chen, J., Liu, X., Tu, P., & Aragones, A. (2013). Learning person-specific models for facial expression and action unit recognition. *Pattern Recognition Letters, 34*, 1964–1970.

Chen, J., Wang, C., Wang, K., Yin, C., Zhao, C., Xu, T., … Yang, T. (2021). HEU Emotion: A large-scale database for multimodal emotion recognition in the wild. *Neural Computing and Applications, 33*, 8669–8685.

Chen, L., Su, W., Feng, Y., Wu, M., She, J., & Hirota, K. (2020). Two-layer fuzzy multiple random forest for speech emotion recognition in human-robot interaction. *Information Sciences, 509*, 150–163.

Cheng, Y., Wang, D., Zhou, P., & Zhang, T. (2018). Model compression and acceleration for deep neural networks: the principles, progress, and challenges. *IEEE Signal Processing Magazine, 35*, 126–136.

Chorowski, J., Bahdanau, D., Cho, K., & Bengio, Y. (2014). End-to-end continuous speech recognition using attention-based recurrent nn: First results. In *NIPS 2014 Workshop on Deep Learning, December 2014*.

Chowdary, M. K., Nguyen, T. N., & Hemanth, D. J. (2021). Deep learning-based facial emotion recognition for human–computer interaction applications. *Neural Computing and Applications*, 1–18.

Chu, W.-S., De la Torre, F., & Cohn, J. F. (2016). Selective transfer machine for personalized facial expression analysis. *IEEE Transactions on Pattern Analysis Machine Intelligence, 39*, 529–545.

Chung, Y.-A., & Glass, J. (2020). Generative pre-training for speech with autoregressive predictive coding. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 3497-3501). Barcelona, Spain: IEEE.

Colnerič, N., & Demšar, J. (2018). Emotion recognition on twitter: Comparative study and training a unison model. *IEEE Transactions on Affective Computing, 11*, 433–446.

D'mello, S. K., & Kory, J. (2015). A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys (CSUR), 47*, 1–36.

Dai, L., Liu, H., Tang, H., Wu, Z., & Song, P. (2022). Ao2-detr: Arbitrary-oriented object detection transformer. *IEEE Transactions on Circuits Systems for Video Technology*.

Dai, W., Cahyawijaya, S., Liu, Z., & Fung, P. (2021a). Multimodal end-to-end sparse model for emotion recognition.

Dai, W., Cahyawijaya, S., Liu, Z., & Fung, P. (2021b). Multimodal End-to-End Sparse Model for Emotion Recognition. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 5305–5316). Mexico City: Association for Computational Linguistics.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V., & Salakhutdinov, R. (2019). Transformer-xl: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 2978–2988). Florence, Italy: Association for Computational Linguistics.

Dang, F., Chen, H., & Zhang, P. (2022). DPT-FSNet: Dual-path transformer based full-band and sub-band fusion network for speech enhancement. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6857-6861): IEEE.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., & Harshman, R. A. (1990). Indexing by latent semantic analysis. *Journal of the Association for Information Science & Technology, 41*, 391–407.

Degottex, G., Kane, J., Drugman, T., Raitio, T., & Scherer, S. (2014). COVAREP—A collaborative voice analysis repository for speech technologies. In *2014 ieee international conference on acoustics, speech and signal processing (ICASSP)* (pp. 960-964). Florence, Italy: IEEE.

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., & Ravi, S. (2020). GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 4040-4054).

Deng, J., & Ren, F. (2021). A survey of textual emotion recognition and its challenges. *IEEE Transactions on Affective Computing*.

Deng, L., & Yu, D. (2014). Deep learning: Methods and applications. *Foundations and Trends in Signal Processing, 7*, 197–387.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics.

Dhall, A., Goecke, R., Lucey, S., & Gedeon, T. (2012). Collecting large, richly annotated facial-expression databases from movies. *IEEE Multimedia, 19*, 34–41.

Eisner, B., Rocktäschel, T., Augenstein, I., Bošnjak, M., & Riedel, S. (2016). emoji2vec: Learning Emoji Representations from their Description. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 48-54). Austin, Texas, USA.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14*, 179–211.

Esperanca, A., & Luo, X. (2022). ReferEmo: A referential quasi-multimodal model for multilabel emotion classification. In C. Strauss, A. Cuzzocrea, G. Kotsis, A. M. Tjoa, & I. Khalil (Eds.), *International Conference on Database and Expert Systems Applications* (pp. 351–366). Cham: Springer International Publishing.

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., … Narayanan, S. S. (2016). The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing, 7*, 190–202.

Eyben, F., Wöllmer, M., & Schuller, B. (2010). Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia* (pp. 1459–1462). Firenze, Italy: ACM.

Fan, X., & Tjahjadi, T. (2015). A spatial-temporal framework based on histogram of gradients and optical flow for facial expression recognition in video sequences. *Pattern Recognition, 48*, 3407–3416.

Fan, Y., Lu, X., Li, D., & Liu, Y. (2016). Video-based emotion recognition using CNN-RNN and C3D hybrid networks. In *Proceedings of the 18th ACM international conference on multimodal interaction* (pp. 445-450). Tokyo, Japan: ACM.

Felbo, B., Mislove, A., Søgaard, A., Rahwan, I., & Lehmann, S. (2017). Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 1615–1625). Copenhagen, Denmark: Association for Computational Linguistics.

Freund, Y., & Haussler, D. (1991). Unsupervised learning of distributions of binary vectors using 2layer networks. *Advances in Neural Information Processing Systems, 4*, 912–919.

Fu, Y., Okada, S., Wang, L., Guo, L., Song, Y., Liu, J., & Dang, J. (2022). Context- and knowledge-aware graph convolutional network for multimodal emotion recognition. *IEEE Multimedia, 29*, 91–100.

García-Martínez, B., Martinez-Rodrigo, A., Alcaraz, R., & Fernández-Caballero, A. (2019). A review on nonlinear methods using electroencephalographic recordings for emotion recognition. *IEEE Transactions on Affective Computing, 12*, 801–820.

Gong, Y., Lai, C.-I., Chung, Y.-A., & Glass, J. (2022). Ssast: Self-supervised audio spectrogram transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, pp. 10699-10709).

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems* (Vol. 27). Montreal, Canada.

Gu, X., Shen, Y., & Xu, J. (2021). Multimodal Emotion Recognition in Deep Learning: a Survey. In *2021 International Conference on Culture-oriented Science & Technology (ICCST)* (pp. 77-82). Beijing, China.

Guo, W., Zhao, X., Zhang, S., & Pan, X. (2022). Learning inter-class optical flow difference using generative adversarial networks for facial expression recognition. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-022-13360-7

Happy, S., & Routray, A. (2017). Fuzzy histogram of optical flow orientations for micro-expression recognition. *IEEE Transactions on Affective Computing, 10*, 394–406.

Hazarika, D., Gorantla, S., Poria, S., & Zimmermann, R. (2018). Self-attentive feature-level fusion for multimodal emotion detection. In *2018 IEEE Conference on*

Multimedia Information Processing and Retrieval (MIPR) (pp. 196–201). IEEE: Miami, FL, USA.
He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).
Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural Computation, 14*, 1771–1800.
Hinton, G. E., Sabour, S., & Frosst, N. (2018). Matrix capsules with EM routing. In *International conference on learning representations (ICLR)*. Vancouver, BC, Canada.
Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science, 313*, 504–507.
Hochreiter, S., & Schmidhuber, J. J. N. c. (1997). Long short-term memory. *9*, 1735-1780.
Hou, M., Zhang, Z., & Lu, G. (2022). Multi-Modal Emotion Recognition with Self-Guided Modality Calibration. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4688-4692). Singapore.
Hu, Y., Zeng, Z., Yin, L., Wei, X., Zhou, X., & Huang, T. S. (2008). Multi-view facial expression recognition. In *2008 8th IEEE International Conference on Automatic Face & Gesture Recognition* (pp. 1-6). Amsterdam, Netherlands.
Huang, J., Tao, J., Liu, B., Lian, Z., & Niu, M. (2020). Multimodal transformer fusion for continuous emotion recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 3507-3511). Barcelona, Spain: IEEE.
Inrak, P., & Sinthupinyo, S. (2010). Applying latent semantic analysis to classify emotions in Thai text. In *2010 2nd International Conference on Computer Engineering and Technology* (Vol. 6, pp. V6-450-V456-454). Chengdu, China: IEEE.
Islam, M. R., Moni, M. A., Islam, M. M., Rashed-Al-Mahfuz, M., Islam, M. S., Hasan, M. K., … Azad, A. (2021). Emotion recognition from EEG signal focusing on deep learning and shallow learning techniques. *IEEE Access, 9*, 94601–94624.
Jahangir, R., Teh, Y. W., Hanif, F., & Mujtaba, G. (2021). Deep learning approaches for speech emotion recognition: State of the art and research challenges. *Multimedia Tools and Applications, 80*, 23745–23812.
Jiang, Y., Li, W., Hossain, M. S., Chen, M., Alelaiwi, A., & Al-Hammadi, M. (2020). A snapshot research and implementation of multimodal information fusion for data-driven emotion recognition. *Information Fusion, 53*, 209–221.
Jung, H., Lee, S., Yim, J., Park, S., & Kim, J. (2015). Joint fine-tuning in deep neural networks for facial expression recognition. In *Proceedings of the IEEE international conference on computer vision* (pp. 2983–2991). Santiago, Chile: IEEE.
Keerthi, S. S., Shevade, S. K., Bhattacharyya, C., & Murthy, K. R. K. (2001). Improvements to Platt's SMO algorithm for SVM classifier design. *Neural Computation, 13*, 637–649.
Keltner, D., Sauter, D., Tracy, J., & Cowen, A. (2019). Emotional expression: Advances in basic emotion theory. *Journal of Nonverbal Behavior, 43*, 133–160.
Khalil, R. A., Jones, E., Babar, M. I., Jan, T., Zafar, M. H., & Alhussain, T. (2019). Speech emotion recognition using deep learning techniques: A review. *IEEE Access, 7*, 117327–117345.
Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2022). Transformers in Vision: A Survey. *54*, Article 200.
Khanpour, H., & Caragea, C. (2018). Fine-grained emotion detection in health-related online posts. In *Proceedings of the 2018 conference on empirical methods in natural language processing (EMNLP)* (pp. 1160–1166). Brussels, Belgium.
Kim, D. H., Baddar, W. J., Jang, J., & Ro, Y. M. (2017). Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition. *IEEE Transactions on Affective Computing, 10*, 223–236.
Krishna, D., & Patil, A. (2020). Multimodal Emotion Recognition Using Cross-Modal Attention and 1D Convolutional Neural Networks. In *Interspeech* (pp. 4243-4247). Shanghai, China: ISCA.
Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems, 25*, 1097–1105.
Kumar, A., & Garg, G. (2019). Sentiment analysis of multimodal twitter data. *Multimedia Tools and Applications, 78*, 24103–24119.
Kumar, A., Srinivasan, K., Cheng, W.-H., & Zomaya, A. Y. (2020). Hybrid context enriched deep learning model for fine-grained sentiment analysis in textual and visual semiotic modality social data. *Information Processing & Management, 57*.
Kwon, S. (2021). MLT-DNet: Speech emotion recognition using 1D dilated CNN based on multi-learning trick approach. *Expert Systems with Applications, 167*, Article 114177.
Lang, P. J. (2005). International affective picture system (IAPS): Affective ratings of pictures and instruction manual. *Technical report*.
Latha, C. P., & Priya, M. (2016). A review on deep learning algorithms for speech and facial emotion recognition. *APTIKOM Journal on Computer Science and Information Technologies, 1*, 92–108.
Latif, S., Rana, R., Khalifa, S., Jurdak, R., Qadir, J., & Schuller, B. W. (2021). Survey of deep representation learning for speech emotion recognition. *IEEE Transactions on Affective Computing*. https://doi.org/10.1109/TAFFC.2021.3114365
Le, D., & Provost, E. M. (2013). Emotion recognition from spontaneous speech using Hidden Markov models with deep belief networks. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding* (pp. 216-221). Olomouc, Czech Republic.
LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*, 436–444.
LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*, 2278–2324.
Li, D., Liu, J., Yang, Y., Hou, F., Song, H., Song, Y., … Mao, Z. (2022). Emotion recognition of subjects with hearing impairment based on fusion of facial expression and EEG topographic map. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*. https://doi.org/10.1109/TNSRE.2022.3225948
Li, S., & Deng, W. (2022). Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing, 13*, 1195–1215.
Li, Y., Tao, J., Schuller, B., Shan, S., Jiang, D., & Jia, J. (2016). MEC 2016: the multimodal emotion recognition challenge of CCPR 2016. In *Chinese Conference on Pattern Recognition* (pp. 667-678). Chengdu, China: Springer.
Li, Y., Zeng, J., Shan, S., & Chen, X. (2019). Occlusion aware facial expression recognition using CNN with attention mechanism. *IEEE Transactions on Image Processing, 28*, 2439–2450.
Lian, Z., Liu, B., & Tao, J. (2021). CTNet: Conversational transformer network for emotion recognition. *IEEE/ACM Transactions on Audio, Speech, Language Processing, 29*, 985–1000.
Liang, J., Chen, S., Zhao, J., Jin, Q., Liu, H., & Lu, L. (2019). Cross-culture Multimodal Emotion Recognition with Adversarial Learning. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4000-4004). Brighton, UK.
Liang, Y., Li, S., Yan, C., Li, M., & Jiang, C. (2021). Explaining the black-box model: A survey of local interpretation methods for deep neural networks. *Neurocomputing, 419*, 168–182.
Lieskovská, E., Jakubec, M., Jarina, R., & Chmulík, M. (2021). A review on speech emotion recognition using deep learning and attention mechanism. *Electronics, 10*, 1163.
Lin, S., Ji, R., Chen, C., Tao, D., & Luo, J. (2019). Holistic CNN compression via low-rank decomposition with knowledge transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41*, 2889–2905.
Liu, J., Chen, S., Wang, L., Liu, Z., Fu, Y., Guo, L., & Dang, J. (2021). Multimodal emotion recognition with capsule graph convolutional based representation fusion. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6339-6343). Toronto, ON, Canada: IEEE.
Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. In *International Conference on Learning Representations (ICLR)* (pp. 1-15).
Luengo, I., Navas, E., & Hernáez, I. (2010). Feature analysis and evaluation for automatic emotion identification in speech. *IEEE Transactions on Multimedia, 12*, 490–501.
Mai, S., Hu, H., Xu, J., & Xing, S. (2022). Multi-fusion residual memory network for multimodal human sentiment comprehension. *IEEE Transactions on Affective Computing, 13*, 320–334.
Mai, S., Zeng, Y., Zheng, S., & Hu, H. (2022). Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. *IEEE Transactions on Affective Computing*. https://doi.org/10.1109/TAFFC.2022.3172360
Mao, Q., Dong, M., Huang, Z., & Zhan, Y. (2014). Learning salient features for speech emotion recognition using convolutional neural networks. *IEEE Transactions on Multimedia, 16*, 2203–2213.
Marín-Morales, J., Llinares, C., Guixeres, J., & Alcañiz, M. (2020). Emotion recognition in immersive virtual reality: From statistics to affective computing. *Sensors, 20*, 5163.
Martin, O., Kotsia, I., Macq, B., & Pitas, I. (2006). The eNTERFACE'05 audio-visual emotion database. In *22nd International Conference on Data Engineering Workshops (ICDEW'06)* (pp. 8-8). Atlanta, GA, USA: IEEE.
Middya, A. I., Nag, B., & Roy, S. (2022). Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities. *Knowledge-Based Systems, 244*, Article 108580.
Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems* (Vol. 2, pp. 3111–3119). Lake Tahoe, Nevada, USA: ACM.
Mittal, T., Bhattacharya, U., Chandra, R., Bera, A., & Manocha, D. (2020). M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 34, pp. 1359-1367).
Narayanan, S., & Georgiou, P. G. (2013). Behavioral signal processing: Deriving human behavioral informatics from speech and language. *Proceedings of the IEEE, 101*, 1203–1233.
Nassif, A. B., Elnagar, A., Shahin, I., & Henno, S. (2021). Deep learning for Arabic subjective sentiment analysis: Challenges and research opportunities. *Applied Soft Computing, 98*, Article 106836.
Niu, Z., Zhong, G., & Yu, H. (2021). A review on the attention mechanism of deep learning. *Neurocomputing, 452*, 48–62.
Noroozi, F., Corneanu, C. A., Kamińska, D., Sapiński, T., Escalera, S., & Anbarjafari, G. (2021). Survey on emotional body gesture recognition. *IEEE Transactions on Affective Computing, 12*, 505–523.
Ottl, S., Amiriparian, S., Gerczuk, M., Karas, V., & Schuller, B. (2020). Group-level speech emotion recognition utilising deep spectrum features. In *Proceedings of the 2020 International Conference on Multimodal Interaction* (pp. 821-826). Utrecht, the Netherlands: ACM.
Ozseven, T. (2023). Infant cry classification by using different deep neural network models and hand-crafted features. *Biomedical Signal Processing and Control, 83*, Article 104648.
Pan, Z., Luo, Z., Yang, J., & Li, H. (2020). Multi-modal attention for speech emotion recognition. In *Interspeech2020*. Shanghai,China: ISCA.
Pandey, S. K., Shekhawat, H. S., & Prasanna, S. M. (2019). Deep learning techniques for speech emotion recognition: A review. In *2019 29th International Conference Radioelektronika* (pp. 1-6). Pardubice, Czech Republic: IEEE.
Peng, S., Cao, L., Zhou, Y., Ouyang, Z., Yang, A., Li, X., … Yu, S. (2021). A survey on deep learning for textual emotion analysis in social networks. *Digital Communications and Networks*. https://doi.org/10.1016/j.dcan.2021.10.003

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543). Doha, Qatar.

Perepelkina, O., Kazimirova, E., & Konstantinova, M. (2018). RAMAS: Russian multimodal corpus of dyadic interaction for affective computing. In *International Conference on Speech and Computer* (pp. 501-510). Leipzig, Germany: Springer.

M.E. Peters M. Neumann M. Iyyer M. Gardner C. Clark K. Lee L. Zettlemoyer Deep contextualized word representations Vol. 1 2018 Association for Computational Linguistics New Orleans, Louisiana 2227 2237.

Poria, S., Chaturvedi, I., Cambria, E., & Hussain, A. (2016). Convolutional MKL based multimodal emotion recognition and sentiment analysis. In *2016 IEEE 16th International Conference on Data Mining (ICDM)* (pp. 439–448).

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., & Mihalcea, R. (2019). MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 527–536). Florence, Italy: Association for Computational Linguistics.

Priyasad, D., Fernando, T., Denman, S., Sridharan, S., & Fookes, C. (2020). Attention driven fusion for multi-modal emotion recognition. In *ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 3227–3231). IEEE: Barcelona, Spain.

Qi, F., Yang, X., & Xu, C. (2021). Zero-shot video emotion recognition via multimodal protagonist-aware transformer network. In *Proceedings of the 29th ACM International Conference on Multimedia* (pp. 1074–1083).

Qian, Y., Zhang, Y., Ma, X., Yu, H., & Peng, L. (2019). EARS: Emotion-aware recommender system based on hybrid information fusion. *Information Fusion, 46*, 141–146.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog, 1*, 9.

Rahman, M. M., Sarkar, A. K., Hossain, M. A., & Moni, M. A. (2022). EEG-based emotion analysis using non-linear features and ensemble learning approaches. *Expert Systems with Applications, 207*, Article 118025.

Ravanelli, M., & Bengio, Y. (2018). Speaker Recognition from Raw Waveform with SincNet. In *2018 IEEE Spoken Language Technology Workshop (SLT)* (pp. 1021-1028). Athens, Greece.

Ren, M., Huang, X., Shi, X., & Nie, W. (2021). Interactive multimodal attention network for emotion recognition in conversation. *IEEE Signal Processing Letters, 28*, 1046–1050.

Rish, I. (2001). An empirical study of the naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence* (Vol. 3, pp. 41-46).

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention* (pp. 234-241): Springer.

Sapiński, T., Kamińska, D., Pelikant, A., Ozcinar, C., Avots, E., & Anbarjafari, G. (2019). Multimodal database of emotional speech, video and gestures. In *International Conference on Pattern Recognition* (pp. 153–163). Beijing, China: Springer International Publishing.

Saxena, D., & Cao, J. (2021). Generative adversarial networks (GANs) challenges, solutions, and future directions. *ACM Computing Surveys (CSUR), 54*, 1–42.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks, 61*, 85–117.

Schmitt, M., Ringeval, F., & Schuller, B. W. (2016). At the border of acoustics and linguistics: Bag-of-audio-words for the recognition of emotions in speech. In *Interspeech* (pp. 495–499). ISCA: San Francisco, USA.

B. Schuller S. Steidl A. Batliner F. Burkhardt L. Devillers C.A. Müller S.S. Narayanan The INTERSPEECH 2010 paralinguistic challenge INTERSPEECH 2010 Makuhari, Chiba, Japan 2794 2797.

Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., Chetouani, M., Weninger, F., Eyben, F., & Marchi, E. (2013). The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. In *INTERSPEECH-2013* (pp. 148-152). Lyon, France.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys (CSUR), 34*, 1–47.

Sebe, N., Cohen, I., Gevers, T., & Huang, T. S. (2005). Multimodal approaches for emotion recognition: a survey. In *Internet Imaging VI* (Vol. 5670, pp. 56-67): SPIE.

Selva, J., Johansen, A. S., Escalera, S., Nasrollahi, K., Moeslund, T. B., Clapés, A. J. I. T. o. P. A., & Intelligence, M. (2023). Video transformers: A survey.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618-626).

Sepas-Moghaddam, A., Etemad, A., Pereira, F., & Correia, P. L. (2020). Facial emotion recognition using light field images with deep attention-based bidirectional LSTM. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 3367–3371). IEEE: Barcelona, Spain.

Shan, C., Gong, S., & McOwan, P. W. (2009). Facial expression recognition based on Local Binary Patterns: A comprehensive study. *Image and Vision Computing, 27*, 803–816.

Sharafi, M., Yazdchi, M., Rasti, R., & Nasimi, F. (2022). A novel spatio-temporal convolutional neural framework for multimodal emotion recognition. *Biomedical Signal Processing and Control, 78*, Article 103970.

Shi, B., Fu, Z., Bing, L., & Lam, W. (2018). Learning domain-sensitive and sentiment-aware word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (Vol. 1, pp. 2494–2504). Melbourne, Australia: Association for Computational Linguistics.

Shoumy, N. J., Ang, L.-M., Seng, K. P., Rahaman, D. M., & Zia, T. (2020). Multimodal big data affective analytics: A comprehensive survey using text, audio, visual and physiological signals. *Journal of Network and Computer Applications, 149*, Article 102447.

Shu, L., Xie, J., Yang, M., Li, Z., Li, Z., Liao, D., … Yang, X. (2018). A review of emotion recognition using physiological signals. *Sensors, 18*, 2074.

Simonyan, K., & Zisserman, A. (2014). *Very deep convolutional networks for large-scale image recognition*. Computer Science.

Soumya George, K., & Joseph, S. (2014). Text classification by augmenting bag of words (BOW) representation with co-occurrence feature. *IOSR Journal of Computing Engineering, 16*, 34–38.

Sun, L., Zhao, G., Zheng, Y., Wu, Z. J. I. T. o. G., & Sensing, R. (2022). Spectral–spatial feature tokenization transformer for hyperspectral image classification. *60*, 1-14.

Sun, N., Li, Q., Huan, R., Liu, J., & Han, G. (2019). Deep spatial-temporal feature fusion for facial expression recognition in static images. *Pattern Recognition Letters, 119*, 49–61.

Sun, Y., Wen, G., & Wang, J. (2015). Weighted spectral features based on local Hu moments for speech emotion recognition. *Biomedical Signal Processing and Control, 18*, 80–90.

Sundberg, J., Patel, S., Bjorkner, E., & Scherer, K. R. (2011). Interdependencies among voice source parameters in emotional speech. *IEEE Transactions on Affective Computing, 2*, 162–174.

Swain, M., Routray, A., & Kabisatpathy, P. (2018). Databases, features and classifiers for speech emotion recognition: A review. *International Journal of Speech Technology, 21*, 93–120.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1-9). Boston, USA.

Tan, Y., Sun, Z., Duan, F., Solé-Casals, J., & Caiafa, C. F. (2021). A multimodal emotion recognition method based on facial expressions and electroencephalography. *Biomedical signal Processing and Control, 70*, Article 103029.

Tan, Y. C., & Celis, L. E. (2019). Assessing social and intersectional biases in contextualized word representations. In *Advances in Neural Information Processing Systems* (Vol. 32, pp. 1-12). Vancouver, BC, Canada.

Taran, S., & Bajaj, V. (2019). Emotion recognition from single-channel EEG signals using a two-stage correlation and instantaneous frequency-based filtering method. *Computer Methods and Programs in Biomedicine, 173*, 157–165.

Ten Bosch, L. (2003). Emotions, speech and the ASR framework. *Speech Communication, 40*, 213–225.

Joseph Raj, A. N., & Gopi, V. P. (2021). Facial Expression Recognition through person-wise regeneration of expressions using Auxiliary Classifier Generative Adversarial Network (AC-GAN) based model. *Journal of Visual Communication and Image Representation, 77*, Article 103110.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017a). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998-6008). Long Beach, CA, USA.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017b). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 6000–6010). Long Beach, CA, USA: ACM.

Vinciarelli, A., Pantic, M., Heylen, D., Pelachaud, C., Poggi, I., D'Errico, F., & Schroeder, M. (2011). Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *IEEE Transactions on Affective Computing, 3*, 69–87.

Wang, D., & Zhao, X. (2022). Affective video recommender systems: A survey. *Frontiers in Neuroscience, 16*, Article 984404.

Wang, Q., Wang, M., Yang, Y., & Zhang, X. (2022). Multi-modal emotion recognition using EEG and speech signals. *Computers in Biology and Medicine, 149*, Article 105907.

Wang, Y., & Guan, L. (2008). Recognizing human emotional state from audiovisual signals. *IEEE Transactions on Multimedia, 10*, 936–946.

Wang, Z., Wan, Z., & Wan, X. (2020). Transmodality: An end2end fusion method with transformer for multimodal sentiment analysis. In *Proceedings of The Web Conference 2020* (pp. 2514-2520). Taipei, Taiwan: Association for Computing Machinery.

Werbos, P. J. (1990). Backpropagation through time: What it does and how to do it. *Proceedings of the IEEE, 78*, 1550–1560.

Windeatt, T. (2006). Accuracy/diversity and ensemble MLP classifier design. *IEEE Transactions on Neural Networks, 17*, 1194–1211.

Wu, C.-H., Lin, J.-C., & Wei, W.-L. (2014). Survey on audiovisual emotion recognition: Databases, features, and data fusion strategies. *APSIPA Transactions on Signal and Information Processing, 3*, 1–12.

Wu, Y., Li, J., Yuan, Y., Qin, A. K., Miao, Q. G., & Gong, M. G. (2022). Commonality autoencoder: Learning common features for change detection from heterogeneous images. *IEEE Trans Neural Netw Learn Syst, 33*, 4257–4270.

Wu, Y., Liu, J.-W., Zhu, C.-Z., Bai, Z.-F., Miao, Q.-G., Ma, W.-P., & Gong, M.-G. (2020). Computational intelligence in remote sensing image registration: A survey. *International Journal of Automation and Computing, 18*, 1–17.

Wu, Y., Xiao, Z., Liu, S., Miao, Q., Ma, W., Gong, M., … Zhang, Y. (2021). A two-step method for remote sensing images registration based on local and global constraints. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14*, 5194–5206.

Xia, M., Wang, K., Song, W., Chen, C., & Li, Y. (2020). Non-intrusive load disaggregation based on composite deep long short-term memory network. *Expert Systems with Applications, 160*, Article 113669.

Xu, B., Fu, Y., Jiang, Y. G., Li, B., & Sigal, L. (2018). Heterogeneous knowledge transfer in video emotion recognition, attribution and summarization. *IEEE Transactions on Affective Computing, 9*, 255–270.

Xu, H., Zhang, H., Han, K., Wang, Y., Peng, Y., & Li, X. (2019). Learning Alignment for Multimodal Emotion Recognition from Speech. In *Proc. Interspeech 2019* (pp. 3569-3573). Graz, Austria.
Yacoub, S. M., Simske, S. J., Lin, X., & Burns, J. (2003). Recognition of emotions in interactive voice response systems. *Interspeech*. Geneva, Switzerland: ISCA.
Yadav, A., & Vishwakarma, D. K. (2020). Sentiment analysis using deep learning architectures: A review. *Artificial Intelligence Review, 53*, 4335–4385.
Yang, K., Wang, C., Gu, Y., Sarsenbayeva, Z., Tag, B., Dingler, T., … Goncalves, J. (2021). Behavioral and physiological signals-based deep multimodal approach for mobile emotion recognition. *IEEE Transactions on Affective Computing*. https://doi.org/10.1109/TAFFC.2021.3100868
Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems, 32*.
Yeasin, M., Bullot, B., & Sharma, R. (2004). From facial expression to level of interest: a spatio-temporal approach. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.* (Vol. 2, pp. II-II). Washington, DC, USA: IEEE.
Yolcu, G., Oztel, I., Kazan, S., Oz, C., Palaniappan, K., Lever, T. E., & Bunyak, F. (2019). Facial expression recognition for monitoring neurological disorders based on convolutional neural network. *Multimedia Tools and Applications, 78*, 31581–31603.
Yu, W., Xu, H., Meng, F., Zhu, Y., Ma, Y., Wu, J., Zou, J., & Yang, K. (2020). Ch-sims: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 3718-3727). Seattle, Washington: Association for Computational Linguistics.
Yu, Y., & Zhang, D. (2022). Few-shot multi-modal sentiment analysis with prompt-based vision-aware language modeling. In *2022 IEEE International Conference on Multimedia and Expo (ICME)* (pp. 1–6). IEEE: Taipei, Taiwan.
Yu, Y., Zhang, D., & Li, S. (2022). Unified Multi-modal Pre-training for Few-shot Sentiment Analysis with Prompt-based Learning. In *Proceedings of the 30th ACM International Conference on Multimedia* (pp. 189-198). Lisboa, Portugal.
Yu, Z., Liu, G., Liu, Q., & Deng, J. (2018). Spatio-temporal convolutional features with nested LSTM for facial expression recognition. *Neurocomputing, 317*, 50–57.
Zadeh, A., Zellers, R., Pincus, E., & Morency, L.-P. (2016). Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems, 31*, 82–88.
Zadeh, A. B., Liang, P. P., Poria, S., Cambria, E., & Morency, L.-P. (2018). 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 (Volume 1: Long Papers)* (pp. 2236-2246). Melbourne, Australia: Association for Computational Linguistics.
Zeng, Z., Pantic, M., Roisman, G. I., & Huang, T. S. (2008). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 31*, 39–58.
Zhalehpour, S., Onder, O., Akhtar, Z., & Erdem, C. E. (2016). BAUM-1: A spontaneous audio-visual face database of affective and mental states. *IEEE Transactions on Affective Computing, 8*, 300–313.
Zhang, J., Yin, Z., Chen, P., & Nichele, S. (2020). Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review. *Information Fusion, 59*, 103–126.
Zhang, S., Liu, R., Tao, X., & Zhao, X. (2021). Deep cross-corpus speech emotion recognition: Recent advances and perspectives. *Frontiers in Neurorobotics, 15*, Article 784514.
Zhang, S., Pan, X., Cui, Y., Zhao, X., & Liu, L. (2019). Learning affective video features for facial expression recognition via hybrid deep learning. *IEEE Access, 7*, 32297–32304.
Zhang, S., Tao, X., Chuang, Y., & Zhao, X. (2021). Learning deep multimodal affective features for spontaneous speech emotion recognition. *Speech Communication, 127*, 73–81.
Zhang, S., Yang, Y., Chen, C., Liu, R., Tao, X., Guo, W., … Zhao, X. (2023). Multimodal emotion recognition based on audio and text by using hybrid attention networks. *Biomedical Signal Processing and Control, 85*, Article 105052.
Zhang, S., Zhang, S., Huang, T., & Gao, W. (2018). Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching. *IEEE Transactions on Multimedia, 20*, 1576–1590.
Zhang, S., Zhang, S., Huang, T., Gao, W., & Tian, Q. (2018). Learning affective features with a hybrid deep model for audio–visual emotion recognition. *IEEE Transactions on Circuits Systems for Video Technology, 28*, 3030–3043.
Zhang, S., Zhang, X., Zhao, X., Fang, J., Liu, M., Zhao, Z., … Tian, Q. (2023). MTDAN: A lightweight multi-scale temporal difference attention networks for automated video depression detection. *IEEE transactions on affective computing*.
Zhang, S., & Zhao, X. (2013). Dimensionality reduction-based spoken emotion recognition. *Multimedia Tools and Applications, 63*, 615–646.
Zhang, S., Zhao, X., & Lei, B. (2012a). Facial expression recognition based on local binary patterns and local fisher discriminant analysis. *WSEAS Transactions on Signal Processing, 8*, 21–31.
Zhang, S., Zhao, X., & Lei, B. (2012b). Robust facial expression recognition via compressive sensing. *Sensors, 12*, 3747–3761.
Zhang, S., Zhao, X., & Tian, Q. (2019). Spontaneous speech emotion recognition using multiscale deep convolutional LSTM. *IEEE Transactions on Affective Computing*.
Zhang, T., Li, S., Chen, B., Yuan, H., & Chen, C. L. P. (2022). AIA-Net: Adaptive interactive attention network for text-audio emotion recognition. *IEEE Transactions on Cybernetics*. https://doi.org/10.1109/TCYB.2022.3195739, 1-13
Zhang, T., Zheng, W., Cui, Z., Zong, Y., Yan, J., & Yan, K. (2016). A deep neural network-driven feature learning method for multi-view facial expression recognition. *IEEE Transactions on Multimedia, 18*, 2528–2536.
Zhang, X., Liu, J., Shen, J., Li, S., Hou, K., Hu, B., … Hu, B. (2021). Emotion recognition from multimodal physiological signals using a regularized deep fusion of kernel machine. *IEEE Transactions on Cybernetics, 51*, 4386–4399.
Zhang, Y., Fu, J., She, D., Zhang, Y., Wang, S., & Yang, J. (2018). Text Emotion Distribution Learning via Multi-Task Convolutional Neural Network. In *International Joint Conference on Artificial Intelligence (IJCAI)* (pp. 4595-4601). Stockholm, Sweden.
Zhang, Z., & Zhang, S. (2023). Modeling air quality PM2. 5 forecasting using deep sparse attention-based transformer networks. *International Journal of Environmental Science Technology*, 1–16.
Zhang, Z., Zhang, S., Zhao, X., Chen, L., & Yao, J. (2022). Temporal difference-based graph transformer networks for air quality PM2. 5 Prediction: A case study in China. *Frontiers in Environmental Science, 10*, Article 924986.
Zhao, G., & Pietikainen, M. (2007). Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 29*, 915–928.
Zhao, J., Li, R., Jin, Q., Wang, X., & Li, H. (2022). Memobert: Pre-Training Model with Prompt-Based Learning for Multimodal Emotion Recognition. In *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4703-4707). Singapore.
Zhao, R., & Mao, K. (2017). Fuzzy bag-of-words model for document representation. *IEEE Transactions on Fuzzy Systems, 26*, 794–804.
Zhao, X., Shi, X., & Zhang, S. (2015). Facial expression recognition via deep learning. *IETE Technical Review, 32*, 347–355.
Zhao, X., & Zhang, S. (2016). A review on facial expression recognition: Feature extraction and classification. *IETE Technical Review, 33*, 505–517.
Zhao, Z., & Liu, Q. (2021). Former-DFER: Dynamic Facial Expression Recognition Transformer. In *Proceedings of the 29th ACM International Conference on Multimedia* (pp. 1553–1561). New York, USA: Association for Computing Machinery.
Zheng, J., Zhang, S., Wang, Z., Wang, X., & Zeng, Z. (2022). Multi-channel weight-sharing autoencoder based on cascade multi-head attention for multimodal emotion recognition. *IEEE Transactions on Multimedia*. https://doi.org/10.1109/TMM.2022.3144885
Zheng, Y. J., Chen, S. B., Ding, C. H. Q., & Luo, B. (2022). Model compression based on differentiable network channel pruning. *IEEE Transactions on Neural Networks and Learning Systems*, 1–10.
Zhou, H., Du, J., Zhang, Y., Wang, Q., Liu, Q.-F., & Lee, C.-H. (2021). Information fusion in attention networks using adaptive and multi-level factorized bilinear pooling for audio-visual emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29*, 2617–2629.
Zia, T., Bashir, N., Ullah, M. A., & Murtaza, S. (2022). SoFTNet: A concept-controlled deep learning architecture for interpretable image classification. *Knowledge-Based Systems, 240*, Article 108066.
Shiqing Zhang, Ruixin Liu, Yijiao Yang, Xiaoming Zhao, Jun Yu. Unsupervised Domain adaptation integrating transformers and mutual information for cross-corpus speech emotion recognition, proceedings of the 30th ACM international conference on multimedia (ACM MM), 120-129. 2022.

Further reading

Chen, B., Cao, Q., Hou, M., Zhang, Z., Lu, G., & Zhang, D. (2021). Multimodal emotion recognition with temporal and semantic consistency. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29*, 3592–3603.
Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N project report, Stanford, 1*, 2009.
Zhao, X., Liao, Y., Xie, J., He, X., Zhang, S., Wang, G., … Yu, J. (2023). BreastDM: A DCE-MRI dataset for breast tumor image segmentation and classification. *Computers in Biology and Medicine, 164*, Article 107255.
Ma, F., Xu, X., Huang, S.-L., & Zhang, L. (2021). Maximum likelihood estimation for multimodal learning with missing modality. *arXiv preprint arXiv:.10513*.
Ma, M., Ren, J., Zhao, L., Tulyakov, S., Wu, C., & Peng, X. (2021). Smil: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 35, pp. 2302–2310).
Poria, S., Cambria, E., Hazarika, D., Majumder, N., Zadeh, A., & Morency, L.-P. (2017). Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 873-883).
Poria, S., Cambria, E., Hazarika, D., Mazumder, N., Zadeh, A., & Morency, L.-P. (2017). Multi-level multiple attentions for contextual multimodal sentiment analysis. In *2017 IEEE International Conference on Data Mining (ICDM)* (pp. 1033-1038). New Orleans, LA, USA: IEEE.
Xu, P., Madotto, A., Wu, C.-S., Park, J. H., & Fung, P. (2018). Emo2vec: Learning generalized emotion representation by multi-task training. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (pp. 292–298). Brussels, Belgium: Association for Computational Linguistics.
Zhao, J., Li, R., & Jin, Q. (2021). Missing modality imagination network for emotion recognition with uncertain missing modalities. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 2608–2618).
Zheng, J., Zhang, S., Wang, X., & Zeng, Z. (2022). Multimodal Representations Learning Based on Mutual Information Maximization and Minimization and Identity Embedding for Multimodal Sentiment Analysis. *arXiv preprint arXiv:2201.03969*.