frontiers | Frontiers in Computer Science

TYPE Original Research
PUBLISHED 09 April 2024
DOI 10.3389/fcomp.2024.1304687

EmoAsst: emotion recognition assistant via text-guided transfer learning on pre-trained visual and acoustic models

Minxiao Wang[1] and Ning Yang[2]*

[1]Computer Engineering Program in School of Electrical, Computer, and Biomedical Engineering, Southern Illinois University, Carbondale, IL, United States, [2]Information Technology Program in School of Computing, Southern Illinois University, Carbondale, IL, United States

OPEN ACCESS

EDITED BY
Andrej Košir,
University of Ljubljana, Slovenia

REVIEWED BY
Hiroki Tanaka,
Nara Institute of Science and Technology (NAIST), Japan
Marko Meža,
University of Ljubljana, Slovenia

*CORRESPONDENCE
Ning Yang
✉ nyang@siu.edu

RECEIVED 29 September 2023
ACCEPTED 05 March 2024
PUBLISHED 09 April 2024

CITATION
Wang M and Yang N (2024) EmoAsst: emotion recognition assistant via text-guided transfer learning on pre-trained visual and acoustic models. *Front. Comput. Sci.* 6:1304687.
doi: 10.3389/fcomp.2024.1304687

Children diagnosed with Autism Spectrum Disorder (ASD) often struggle to grasp social conventions and promptly recognize others' emotions. Recent advancements in the application of deep learning (DL) to emotion recognition are solidifying the role of AI-powered assistive technology in supporting autistic children. However, the cost of collecting and annotating large-scale high-quality human emotion data and the phenomenon of unbalanced performance on different modalities of data challenge DL-based emotion recognition. In response to these challenges, this paper explores transfer learning, wherein large pre-trained models like Contrastive Language-Image Pre-training (CLIP) and wav2vec 2.0 are fine-tuned to improve audio- and video-based emotion recognition with text- based guidance. In this work, we propose the EmoAsst framework, which includes a visual fusion module and emotion prompt fine-tuning for CLIP, in addition to leveraging CLIP's text encoder and supervised contrastive learning for audio-based emotion recognition on the wav2vec 2.0 model. In addition, a joint few-shot emotion classifier enhances the accuracy and offers great adaptability for real-world applications. The evaluation results on the MELD dataset highlight the outstanding performance of our methods, surpassing the majority of existing video and audio-based approaches. Notably, our research demonstrates the promising potential of the proposed text-based guidance techniques for improving video and audio-based Emotion Recognition and Classification (ERC).

KEYWORDS

emotion recognition, transfer learning, pre-trained model, contrastive learning, multi-modal

1 Introduction

Many children who are diagnosed with Autism Spectrum Disorder (ASD) have difficulty understanding social conventions and identifying sarcasm, humor, or figurative language. They struggle with conversational turn-taking and interpreting social cues, making it difficult to initiate and maintain friendships and relationships. These difficulties may manifest as tantrums, anxiety, aggressive behavior, and a tendency to become easily frustrated. The occurrence of inappropriate behavior can be attributed to various factors related to physical and psychological aspects. A primary factor is their difficulty in promptly recognizing others' emotions and accurately interpreting facial expressions, leading to challenges in adjusting to appropriate responses. This motivated us to use machine learning technology to help autistic children better recognize people's emotions. AI-powered assistive technology refers to the use of artificial intelligence (AI) to develop tools and devices that assist individuals with disabilities in various aspects of their daily lives. These technologies are designed to enhance independence, accessibility, and overall quality of life for people with disabilities. Recent advances in utilizing deep learning for recognizing emotions are establishing it as a valuable assistive technology, capable of providing substantial support to individuals across multiple applications and domains.

However, collecting and annotating large amounts of high-quality data is costly and even impossible for some special domains, such as emotion detection. Meanwhile, existing AI-based emotion recognition methods have imbalanced performance on different independent modalities (audio, video, and text). For example, on most existing benchmarks (Busso et al., 2008; Poria et al., 2018), text-based emotion recognition methods always achieve much better detection accuracy than audio or video-based methods. Additionally, it should be noted that the text of speech is not intuitive sensory information like the acoustic and visual information in the formats of audio and video. Furthermore, for the aspect of improving emotion recognition abilities, acoustic and visual information are more useful (Ghaleb et al., 2019; Ma et al., 2020) than texts. Therefore, we have considered how to improve the audio and video-based emotion recognition performance based on the guidance from successful text-based methods.

Recently, significant progress has been made in visual representation learning through large-scale contrastive vision-language pre-training (CLIP) (Radford et al., 2021). CLIP is a deep learning model developed by OpenAI that learns visual concepts by training on a large dataset of images paired with natural language descriptions. The language knowledge learned in CLIP helps the model understand the semantics or meanings associated with various concepts in images. Hence, it can understand and represent images in a way that is useful for a wide range of tasks. Although CLIP initially focused on images and text, it has been extended and adapted to learn from audio and video data. This cross-modal learning and generalization enable a broader understanding of multimodal information and facilitate various applications in the audio and video domains (Xu et al., 2021; Ma et al., 2022; Zhang et al., 2023). Furthermore, many recent works focus on adapting the pre-trained CLIP models for various downstream applications (Zhang et al., 2021; Lin et al., 2022; Rasheed et al., 2023).

In order to solve the data collection and performance imbalance issues for deep learning-based emotion recognition, we believe that fine-tuning the pre-trained foundation models offers a good solution. In this paper, we propose transfer learning methods to take advantage of large pre-trained models, particularly CLIP and wav2vec 2.0, for video and audio-based emotion ERC tasks. We improve the audio and video-based emotion recognition performance based on the guidance from text-based methods. The main contributions are listed below:

- Design a visual fusion module and an emotion prompt fine-tuning method to improve visual emotion representations of CLIP with the guidance of texts.
- Adopt CLIP's text encoder and use supervised contrastive learning to improve transfer learning on another pre-trained model (wav2vec 2.0) for audio-based ERC tasks.
- Design a joint few-shot emotion classifier for the fine-tuned visual and acoustic representations to achieve better accuracy on the video and audio-based ERC.

Evaluation results on the MELD dataset showed that our methods outperformed existing video and audio-based methods and all of the proposed methods can bring benefits to video and audio-based ERC.

The remained of this presents, related work in Section 2. Section 3 outlines our methodology. Section 4 analyzes the experiments and the evaluation results. Conclusions are described in Section 5 and Section 6 discusses future work.

2 Related work

Previous research studies have established connections between challenges in narrative skills in individuals with ASD to deficits in social cognitive abilities. These challenges can include difficulties in accurately interpreting the emotions and cognitive states of others, which could potentially restrict their ability to respond appropriately in social situations (Tager-Flusberg, 2000; Losh and Capps, 2006). An important factor contributing to social difficulties in children with ASD is emotion recognition, which involves the ability to accurately identify and interpret emotions based on facial expressions, vocal cues, body language, and contextual information. Assistive technology can greatly assist individuals by providing a supportive and engaging environment for learning and practicing emotion recognition in a structured and effective manner. The use of machine learning in emotion recognition can analyze vast amounts of related data, and enable the development of personalized solutions based on individual needs, preferences, and abilities that meet children's unique social challenges.

2.1 Emotion recognition

Emotion recognition research is a continuously evolving and dynamic domain. Researchers continue to explore various approaches and technologies to improve the accuracy and applicability of emotion recognition systems. Deep learning models, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and more recently, Transformer-based models, have shown significant promise in emotion recognition tasks (Kahou et al., 2015; Fan et al., 2016; John and Kawanishi, 2022; Febrian et al., 2023). Transfer learning, in which pre-trained models are fine-tuned for emotion recognition tasks, has gained popularity due to its ability to leverage large-scale labeled datasets (Feng and Chaspari, 2020).

Over the past few years, there has been an increasing focus on multimodal emotion recognition using deep learning and signal processing methodologies (Tashu et al., 2021; Ma et al., 2022). Combining information from multiple modalities (e.g., facial expressions, speech, text, physiological signals) has been a focus to improve the accuracy and robustness of emotion recognition. Integrating data from different modalities (e.g., text, audio, video) has shown improved performance compared to using a single modality. However, each modality has its own features, structure, and noise, and extracting meaningful features from diverse modalities and fusing them into a cohesive representation that captures emotional content is a key challenge (Zhang et al., 2018).

Emotion recognition in real-world scenarios often requires considering the contextual and temporal aspects of emotional expressions. Contextual information from conversations, interactions, or surrounding events can help to better understand and interpret emotions. Children with ASD may exhibit slower responses to people or social cues. Integrating emotion recognition tools into educational and therapeutic interventions can offer valuable support to enhance their understanding of emotions in various social contexts.

2.2 Pre-trained models

Pre-training models on large-scale text data help the model learn rich and abstract representations of language. These learned features can be transferable and beneficial for understanding text in various domains, including emotion recognition. Considering the efficient use of data, pre-training leverages vast amounts of readily available unlabeled text data. This enables the model to learn from diverse and extensive linguistic patterns and nuances without requiring large amounts of labeled emotion-specific data. Pre- training also facilitates transfer learning, allowing the model to use its learned knowledge and representations from a general task such as language understanding, and apply them to a specific task like emotion recognition, reducing the amount of labeled data needed for the specific task.

Contrastive Language-Image Pre-training (CLIP), created by OpenAI, is an advanced deep-learning model that matches natural language descriptions with images, enabling a broad spectrum of vision-related tasks. It simultaneously trains an image encoder and a text encoder to correctly associate pairs of (image, text) training examples within a batch. During testing, the text encoder generates a zero-shot linear classifier by embedding the names or descriptions of the classes in the target dataset (Radford et al., 2021). CLIP is a flexible approach that can be applied to various image understanding and processing tasks without the need for task-specific training. It can recognize broad categories and concepts in images, but may struggle with fine-grained object recognition or distinguishing subtle differences within similar categories. Furthermore, it has limited contextual understanding because it operates on an image-by-image basis without considering contextual information or relationships between multiple objects or entities within an image. It may not capture complex spatial or contextual dependencies.

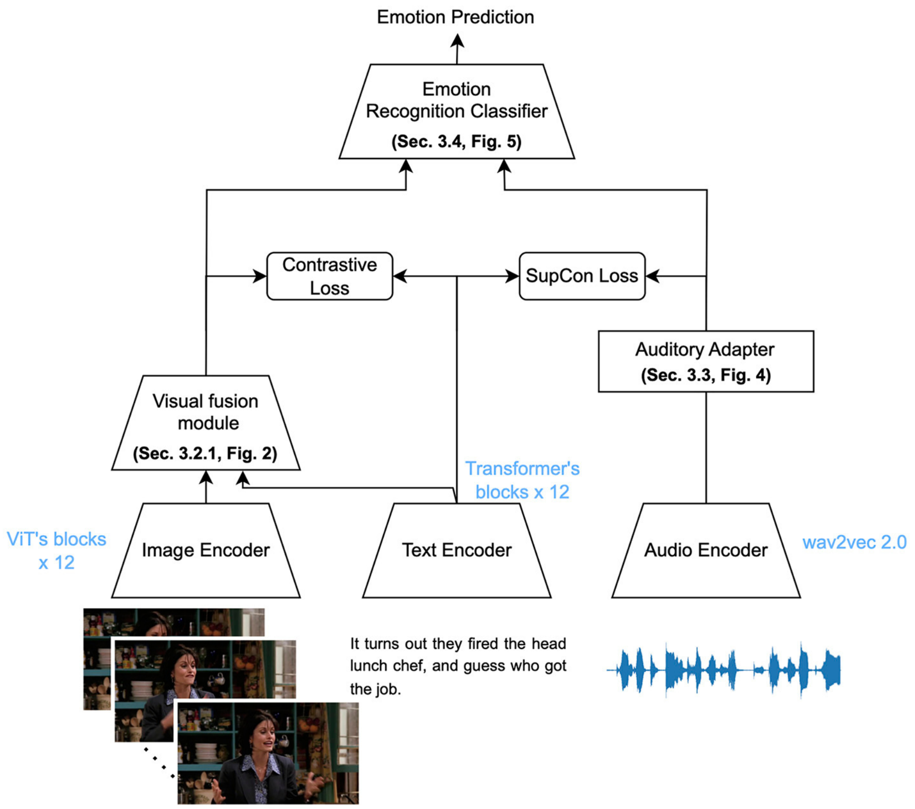


FIGURE 1
Overview of the EmoAsst framework. EmoAsst comprises three pre-trained encoders, all of which are frozen during operation. The image and text encoders are sourced from the CLIP model, while the audio encoder is based on the wav2vec 2.0 model. A visual adapter (visual fusion module) and an auditory adapter are proposed for transferring the pre-trained knowledge for ERC downstream tasks with contrastive learning. A cache model-based emotion recognition classifier is proposed for final prediction.

Facebook AI Research has developed a cutting-edge framework for self-supervised learning of speech representations, Wav2Vec 2.0 (Baevski et al., 2020a). The model is designed to transform raw audio signals into a more abstract and informative representation that can be used for downstream speech-related tasks like transcription or translation. Wav2Vec 2.0 employs a self-supervised learning approach, which allows it to be pre-trained on a large amount of unlabeled data. The key advantage lies in its architecture and training methodology, particularly in its audio encoder and subsequent use for downstream audio decoding.

3 Methods

In this study, we transferred and combined the pre-trained CLIP and wav2vec 2.0 models to the downstream emotion recognition task. In particular, we focused on improving the performance of video and audio-based emotion recognition based on the guidance of learned representation features from text-based methods.

To this end, we proposed a CLIP-based multi-modal contrastive transfer learning framework (Section 3.1). Particularly, we first proposed a visual emotion representation fine-tuning method based on the pre-trained CLIP model in Section 3.2; then, we proposed an acoustic emotion representation transfer learning method with pre-trained features from the wav2vec 2.0 model and fine-tuned with the CLIP structure in Section 3.3; finally, we proposed a multi-modal emotion recognition classifier for the transferred visual and acoustic emotion representation features in Section 3.4.

3.1 Framework overview

An overview of our transfer learning framework is presented in Figure 1. The EmoAsst framework includes (1) the most widely adopted vision-language contrastive learning paradigm, CLIP, which consists of an image encoder and a text encoder for extracting visual and language features; (2) a wav2vec 2.0 model, which works as an audio encoder; and (3) an emotion recognition classifier for predicting emotion with the transfer emotion representation feature.

In EmoAsst framework, we adopted the pre-trained CLIP and wav2vec 2.0 models with frozen weights. In order to adapt the pre-trained CLIP to a video-based emotion recognition downstream task, we first bridged the gap between images and videos by appending a visual fusion module after the image encoder. Next, we added a visual adapter to transfer the learned visual representations to the emotion-related domain. Finally, we used a text prompting method to train the visual fusion module and visual adapter with contrastive loss. To adapt the pre-trained wav2vec 2.0 to audio- based emotion recognition downstream tasks, we combined wav2vec within the CLIP model as a triple-branch CLIP. As an audio encoder, wav2vec 2.0 extracts acoustic representations that are then transferred to the emotion-related domain by the following auditory adapter. The audio branch is also fine-tuned with the same text prompting method in the video branch.

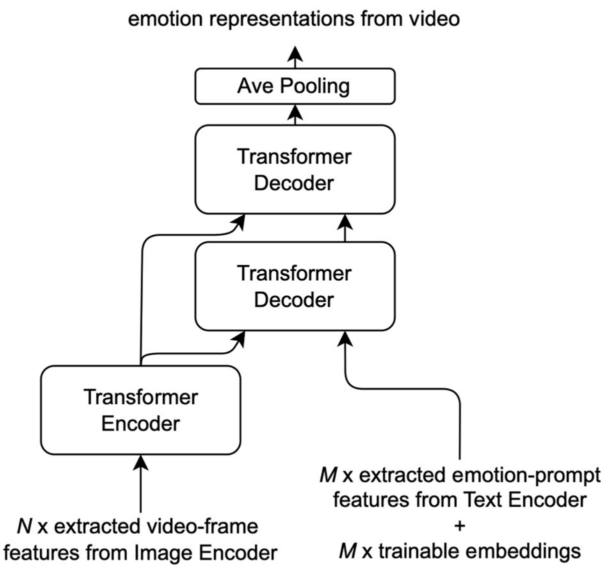

emotion representations from video
Ave Pooling
Transformer Decoder
Transformer Decoder
Transformer Encoder
N x extracted video-frame features from Image Encoder
M x extracted emotion-prompt features from Text Encoder + *M* x trainable embeddings

FIGURE 2
The structure of the visual fusion module.

3.2 Visual emotion representation fine-tuning on CLIP

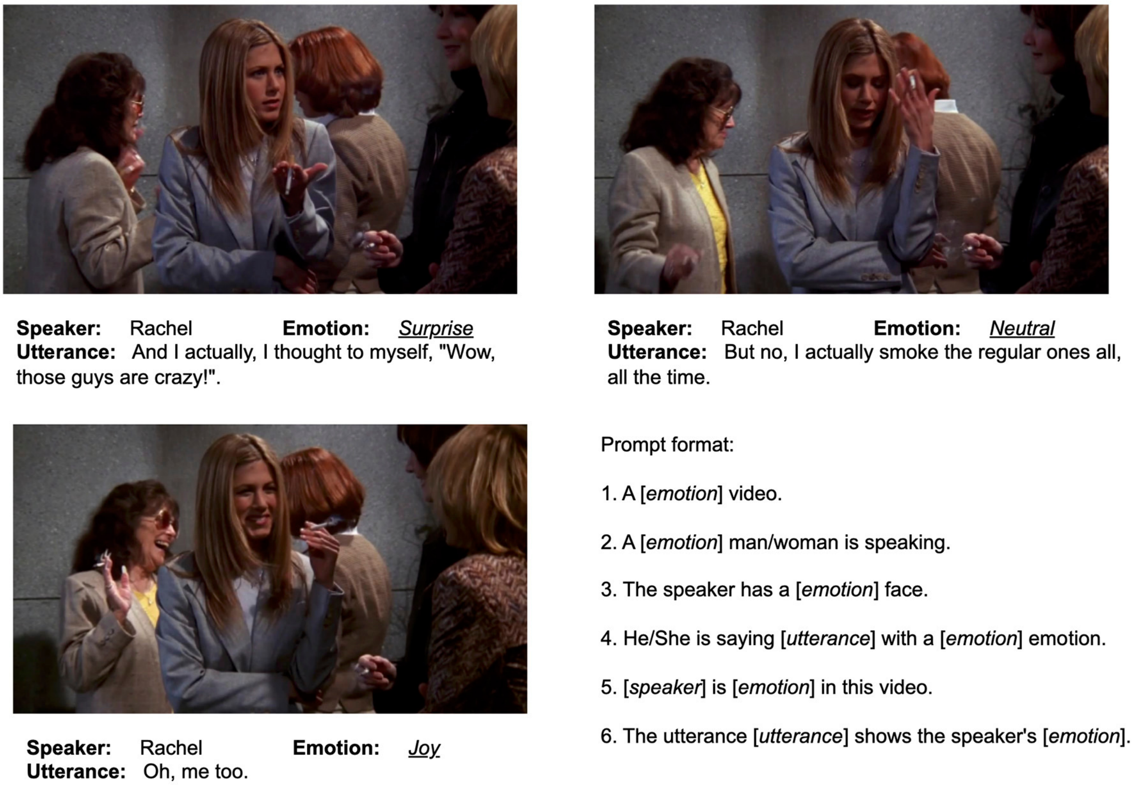
3.2.1 Adapting image CLIP for visual emotion

Many existing works (Lin et al., 2022; Rasheed et al., 2023) report that large-scale image-text-based pre-trained CLIP can be transferred to the video learning domain by adding extra learnable parametric modules (normally stacked LSTM layers or transformer encoders) to model the temporal relation among frames. However, based on the analysis of EmotionCLIP (Zhang et al., 2023), the transferred video-level CLIP for the visual emotion recognition task cannot achieve comparable success in other video- based tasks, such as human action understanding. We believe the main reason is that emotional expressions may not always be easily discernible. In a video, the primary focus lies in the subject's movements and actions, with their internal emotions often taking a secondary role as supplementary information. Therefore, we designed a novel visual fusion module to bridge the modality gap between images and video.

3.2.1.1 Visual fusion module

Instead of modeling the temporal or context relationships in videos, we used the image fusion module to extract consistent emotional expressions. As shown in Figure 2, unlike the existing transformer encoder-based temporal video fusion modules (Ma et al., 2022; Zhang et al., 2023), our visual fusion module consists of one transformer encoder layer and two transformer decoder layers. The transformer encoder works as an adapter to fine-tune the representation of each individual video frame independently. The sequence of fine-tuned representations of N video frames is further fed to the transformer decoder layers.

The transformer decoders also take M query features as an additional input. The query feature is the summation of a trainable embedding vector and an emotion-prompt feature vector extracted



Speaker: Rachel **Emotion:** *Surprise*
Utterance: And I actually, I thought to myself, "Wow, those guys are crazy!".

Speaker: Rachel **Emotion:** *Neutral*
Utterance: But no, I actually smoke the regular ones all, all the time.

Speaker: Rachel **Emotion:** *Joy*
Utterance: Oh, me too.

Prompt format:

1. A [*emotion*] video.
2. A [*emotion*] man/woman is speaking.
3. The speaker has a [*emotion*] face.
4. He/She is saying [*utterance*] with a [*emotion*] emotion.
5. [*speaker*] is [*emotion*] in this video.
6. The utterance [*utterance*] shows the speaker's [*emotion*].

FIGURE 3
The character Rachel had three different emotion types during the smoking action in the same scenario. The similar temporal and contextual information makes it hard to recognize the different emotions. To solve this, we proposed an emotion prompt to fine-tune the emotion representation features. Six examples of prompt format are given.

from the text encoder. The emotion-prompt features are used to calculate cross-attention in the transformer decoder layers with the sequence of fine-tuned representations. Details about emotional prompting will be introduced in Section 3.2.2. Due to the participation of emotion-prompt features, the transformer decoder layers can extract the emotion-related features from the sequence of N fine-tuned representations of video frames. The added embedding vector makes the extraction process trainable instead of using the fixed emotion-prompt features.

3.2.2 Fine-tuning the emotion prompt

In this section, we introduce the adopted emotion prompt engineering technique for fine-tuning the pre-trained CLIP and wav2vec 2.0 model. Although CLIP has also been adopted for learning emotion representations by EmotionCLIP (Zhang et al., 2023), emotion-related prompting has not been included in those works. This is because the research scope of our work is different from EmotionCLIP. EmotionCLIP is a pre-training framework that which can only use uncurated data from the internet, such as YouTube, to learn emotion representations based on video and text communication information. As a pre-training paradigm, EmotionCLIP only needs to provide a few sentiment cues to force the large model to learn emotion-related knowledge. But our EmoAsst focuses on a specific downstream task—emotion recognition, we can provide more conspicuous emotion-related guidance in our approach.

As shown in Figure 3, the three pictures are three video frames that come with three videos with different emotion labels in the MELD dataset. However, those three videos have the same speaker (Rachel), the same background, and the same action (smoking). In this case, we should notice that only capturing the temporal and contextual relationships may be enough for the action recognition task but not enough to distinguish different emotions. This scenario is quite common in real-life conversations. In daily conversations, the subjects of the conversation, as well as the scene and activities they are engaged in usually remain the same.

To fine-tune the visual emotion features, we customized the prompt text to emphasize the emotion-related context. In EmoAsst, we added the emotion prompts to the text information, which is the input of the CLIP text encoder. In our approach, we adopted six prompt formats to augment the text representations so that the text features can help fine-tune the visual features to include more emotion-related information. In our prompt format, the emotion label [*emotion*] must be included. Meanwhile, to emphasize the emotion, we used some keywords, such as "speaking", "say", and "face" to guide the visual fusion module to focus more on the speaker's face. To increase the diversity of prompts, [*speaker*] and [*utterance*] information were optionally added to the prompt sentences.

Another compelling rationale for employing multiple prompt formats lies in their capacity to mitigate the issue of false negative pairs. In the context of the CLIP model's contrastive loss, all video-text intersection pairs are mandated to be categorized as negative.

Nevertheless, it is untenable to enforce a negative label on video-text pairs that share the same emotion label. To address this concern, we introduced a variety of prompt texts, complemented by additional contextual information such as utterance content and speaker name or gender.

3.3 Fine-tuning acoustic emotion representation in CLIP and wav2vec 2.0

Since the original CLIP does not include an audio encoder, we adopted the wav2vec 2.0 model as the audio encoder for EmoAsst. Although wav2vec shares a similar transformer structure with the text encoder of CLIP, the large training cost makes it impossible to pre-train wav2vec 2.0 from scratch. Additionally, in this work, we focused on text-guided transfer learning. Therefore, we used a pre-trained wav2vec 2.0 base encoder whose checkpoint was trained by SEGUE (Tan et al., 2023). SEGUE is a pre-training method for spoken language understanding (SLU) tasks. Although SEGUE did not use the text encoder from CLIP, it did use wav2vec 2.0 as its audio encoder. Furthermore, SEGUE's pre-trained audio encoder was evaluated on the same MELD (Poria et al., 2018) dataset as our work, which proves it can provide a good starting point for our task.

3.3.1 Model structure

Specifically, we adopted the frozen CTC-based feature encoder and the frozen first 9 transformer layers of the context network of the wav2vec 2.0 structure, followed by three extra trainable transformer layers, called auditory adapter (AA), to match the different feature dimensions (768-dimensional embedding for wav2vec 2.0 and 512- dimensional embedding for CLIP) and the different downstream tasks of ASR and emotion recognition. The structure of AA is shown in Figure 4 AA projects the extracted representation of the pre-trained audio encoder into the same representation space as the text encoder of CLIP.

3.3.2 Training objective

During the training of the audio encoder, two levels of transfer learning are included: (1) cross-task transfer, where the auditory adapter is trained to transfer the extracted features from the SLU task to the emotion recognition task; (2) cross-modality transfer, where the CLIP text encoder works as a teacher network and the frozen wav2vec 2.0 with auditory adapter is a student network. Although the SEGUE pre-training bridged the modality gap to a certain extent, the difference in the dimensions of the representation feature reduced its reusability.

To efficiently train the AA structure for the downstream emotion recognition task, we adopted the supervised contrastive (SupCon) loss (Khosla et al., 2020), which is defined as Equation 1:

$$\mathcal{L} = \sum_{i \in M} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(z_i \cdot z_p/\tau)}{\sum_{a \in A(i)} \exp(z_i \cdot z_a/\tau)} \tag{1}$$

where i is the data sample index of the set M, consisting of both the extracted audio and text features z_i, so the size of M is two times the batch size. The $P(i)$ is the set of samples that have the same class as sample i. The $A(i)$ is the set of all the other samples except i.

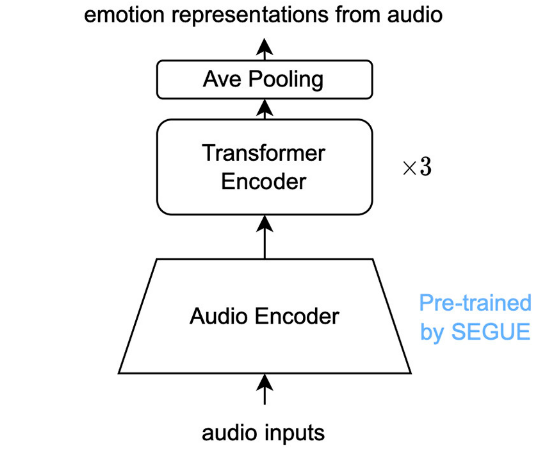


FIGURE 4
The structure of the auditory adapter.

With the guidance of the labels, the positive/negative samples are clearly defined. Compared with the contrastive loss used in CLIP, the SupCon loss also considers the contrast within the same modality representations, which can help to distinguish the acoustic representations with different emotions. Furthermore, the supervised loss can encourage the audio encoder to learn closely aligned representations of all samples from the same class.

3.4 Emotion recognition classifier

In this work, we aimed to apply the transferred pre-trained CLIP and wav2vec 2.0 for emotion recognition assistant technology. The zero-shot performance cannot meet the recognition accuracy requirement. To further improve the recognition performance, we adopt a cache model-based emotion recognition classifier, which is inspired by Tip-Adapter. Zhang et al. (2021) show that Tip- Adapter can achieve comparable performance compared to those CLIP adaptation methods that require training without fine-tuning. Additionally, it can be further improved by few-shot fine-tuning on the cache model.

As shown in Figure 5, the emotion recognition classification joins the zero-shot classifier and the joint cache model. The zero-shot classifier is constructed from the extracted features generated by the text encoder of EmoAsst. For all N emotion labels, text prompts are fed into the text encoder to generate F_{cls}^{text}, which are the weights of the zero-shot classifier. The zero-shot classification is calculated as

$$pred_{zero} = (f_{test}^{video} \mathrm{F}_{\mathrm{cls}}^{\mathrm{text}^{\boldsymbol{T}}} + f_{test}^{audio} \mathrm{F}_{\mathrm{cls}}^{\mathrm{text}^{\boldsymbol{T}}})/2, \tag{2}$$

where $f_{test}^{video}, f_{test}^{audio} \in \mathbb{R}^{1\times d}$, and $\mathrm{F}_{\mathrm{cls}}^{\mathrm{text}^{\boldsymbol{T}}} \in \mathbb{R}^{N\times d}$.

To improve the zero-shot emotion recognition classification in Equation 2, a K-shot joint cache model is introduced. As shown in the left part of Figure 5, K-shot training data are collected and fed

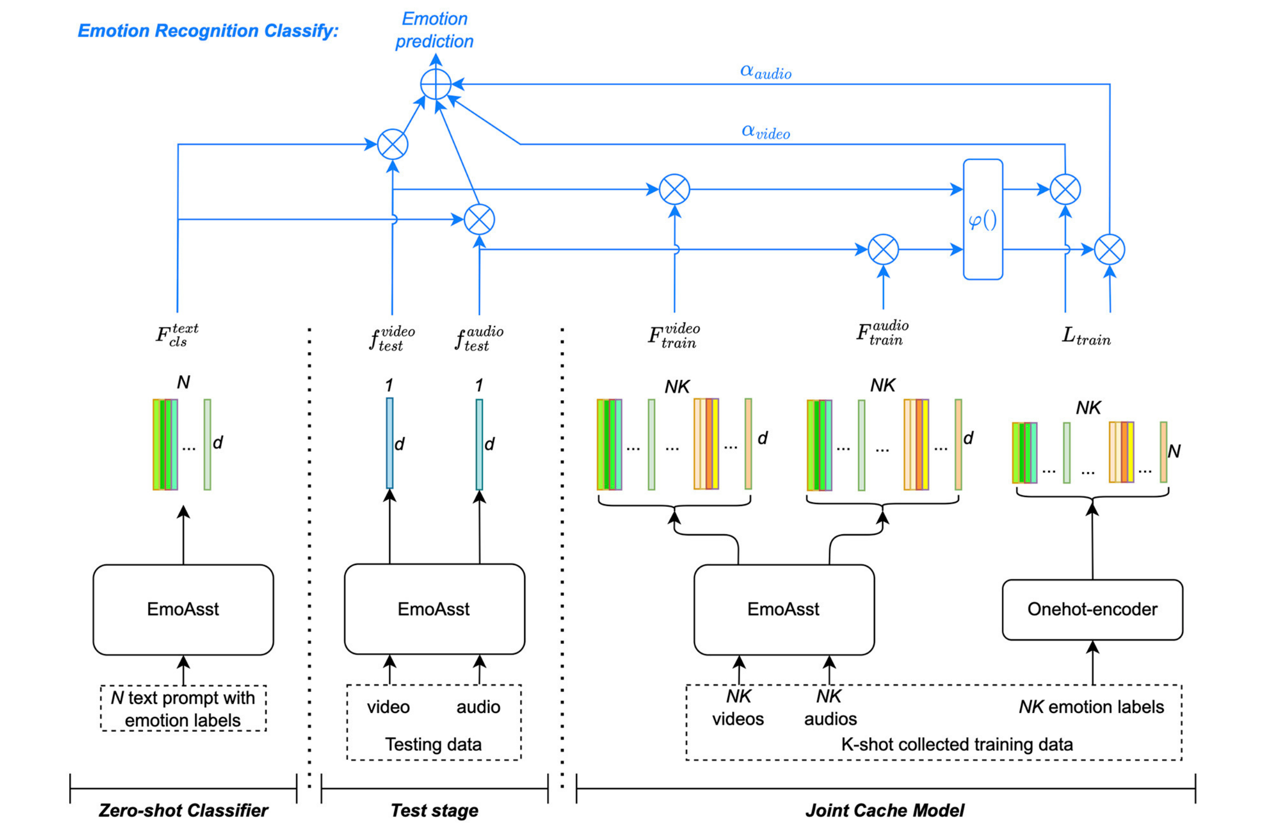


FIGURE 5
The classification process is shown in this figure. The joint cache model is depicted on the left side. Here, the K-shot visual and acoustic representations serve as the keys within the cache model, while their respective labels are one-hot encoded as values. During the testing phase, the extracted visual and acoustic features are treated as queries, and information is gathered from the cache model through similarity-based retrieval. This process is employed to calculate the probability of emotion recognition.

into the fine-tuned EmoAsst to extract visual and acoustic emotion representations as the *Keys* of the cache model, denoted as $\mathrm{F}_{\mathrm{train}}^{\mathrm{video}}$ and $\mathrm{F}_{\mathrm{train}}^{\mathrm{audio}}$. The corresponding labels are encoded into one-hot vectors as the *Values* of the cache model, denoted as $\mathrm{L}_{\boldsymbol{train}}$. The K-shot classification is calculated in Equation 3 as:

$$pred_k = \alpha_{video}\varphi(f_{test}^{video}{\mathrm{F}_{\mathrm{train}}^{\mathrm{video}}}^{\boldsymbol{T}})\mathrm{L}_{\mathrm{train}} + \alpha_{audio}\varphi(f_{test}^{audo}{\mathrm{F}_{\mathrm{train}}^{\mathrm{audio}}}^{\boldsymbol{T}})\mathrm{L}_{\mathrm{train}} \tag{3}$$

where ${\mathrm{F}_{\mathrm{train}}^{\mathrm{video}}}^{\boldsymbol{T}}$, ${\mathrm{F}_{\mathrm{train}}^{\mathrm{audio}}}^{\boldsymbol{T}} \in \mathbb{R}^{NK\times d}$ and $\mathrm{L}_{\mathrm{train}} \in \mathbb{R}^{NK\times N}$. α is the residual ratio, and the function $\varphi()$ is defined as Equation 4:

$$\varphi(x) = exp(-\beta(1 - f_{test}^{video}{\mathrm{F}_{\mathrm{train}}^{\mathrm{video}}}^{\boldsymbol{T}})), \tag{4}$$

where β represents a modulating hyper-parameter that is independent of the video or audio features. The final prediction of the emotion recognition classifier is $pred_{zero} + pred_k$, which is the combination of Equations 2, 3.

Because the joint prediction of two modalities, video and audio, is a weighted summation, the emotion recognition classifier can work for both single modality and multiple modalities by simply choosing the workflow pipeline. The single modality classifier is used for the evaluation in Sections 4.2 and 4.3. The training-free cache model classifier can also be easily updated in real-world application scenarios by collecting K-shot training data from the target application scenarios.

4 Results and discussion

In this section, first, we describe the experimental settings, including datasets, training configurations, and baseline models in Section 4.1. Then we present the performance improvement on video-based and audio-based emotion recognition separately with guidance from language-based methods in Sections 4.2 and 4.3.

4.1 Dataset and experimental settings

4.1.1 Dataset

We leveraged the Multimodal EmotionLines Dataset (MELD) (Poria et al., 2018) to use information from multiple parallel data channels, including video, audio, and text. MELD was developed by expanding upon and enriching the original EmotionLines dataset (Chen et al., 2018). MELD boasts a comprehensive collection of over 1,400 dialogues and 13,000 utterances sourced from the Friends television series, featuring contributions from multiple speakers. MELD includes seven emotions for the annotation, which are anger, disgust, fear, joy, neutral, sadness, and surprise, across the training, validation, and testing splits. More statistical information is presented in Table 1. Compared with other multimodal emotion

TABLE 1 The statistical information of the MELD dataset.

Statistic	Train	Dev	Test
Num of modality	3	3	3
Num of dialogues	1,039	114	280
Num of utterances	9,989	1,109	2,610
Avg utterance length	8.03	7.99	8.28
Max utterance length	69	37	45
Avg duration of an utterance	3.59s	3.59s	3.58s

recognition datasets, MELD presents conversational scenarios that are more similar to the real-life application scenarios of emotion recognition assistance technology. In this work, we used the official training, validation, and test split. A limitation of this work is that the experiment only evaluated one dataset. The reason we only used one dataset is that not many datasets include both text and video modalities. In addition to the MELD dataset used in this work, other datasets, such as IEMOCAP, normally were generated in a lab environment. However, in this work, we needed a dataset for the conversation in real-life scenes. Therefore, other existing datasets were not suitable for our goal.

4.1.2 Experimental settings

Our EmoAsst framework and the pre-trained models were implemented in Pytorch (Paszke et al., 2019). We fine-tuned our EmoAsst framework on an NVIDIA GeForce RTX 4070 Ti GPU with a batch size of 128. For the optimizer, we adopted AdamW (Loshchilov and Hutter, 2017) with a 0.01 scale of weight decay regularization and a group of learning rates, 0.001 for the Visual Fusion Module, and 0.0001 for the auditory adapter. As we mentioned in Section 3.2.1, VFM includes a transformer encoder and two transformer decoders, where the encoder has 512 input dimensions for self-attention layers and 2,048 dimensions for feed–forward networks, while the decoder has 512 for both text embedding and frame features for self-attention layers, and 2,048 dimensions for feed-forward networks. The total VFM parameter count is 11.56 M. The AA, in Section 3.3, includes three transformer encoders that have 512 input dimensions for self-attention layers, and 2048 dimensions for feed-forward networks. The total AA parameter count is 9.46 M. Pre-training the VFM and AA on the training set in Table 1 required 200 epochs (21.22 min per epoch) to achieve the results shown in Sections 4.2 and 4.3.

For each text input, the prompt format was randomly chosen from the six given examples for the training stage. During the testing stage, as the emotion labels are not participating in the inference, the text input will only use the utterance text. For the input video clips, the number of sparsely sampled frames was 24 for each clip. Due to the video clips in the MELD dataset having various lengths, the implementation had to choose a fixed sampling number instead of a sampling ratio to ensure the input shape.

To evaluate our method, we adopted recognition accuracy and weighted F1 score as evaluation metrics. The weighted F1 score took into account the imbalance in class ratio. Each class was assigned a weight based on its relative proportion in the dataset. The weighted F1 score is calculated with Equation 5:

TABLE 2 The performance of fine-tuned visual emotion representations compared with other existing video-level CLIP in terms of accuracy and weighted F1 score by linear-probe evaluation.

Method	Linear-Eval	
	Acc	Weighted F1
VideoCLIP (Xu et al., 2021)	45.19	32.06
X-CLIP (Ma et al., 2022)	38.31	32.46
EmotionCLIP (Zhang et al., 2023)	48.28	34.59
VFM	49.36	36.31
VFM + prompt	**50.37** + 1.01	**37.13** + 0.82

The bold value is the highest value.

$$WF1 = \frac{\sum_{i=1}^{N} w_i \times F1_i}{\sum_{i=1}^{N} w_i} \tag{5}$$

4.1.3 Pre-trained baseline

We adopted the pre-trained EmotionCLIP model as our visual baseline model. The pre-trained EmotionCLIP applied a ViT (Dosovitskiy et al., 2020) with a patch size of 32 and an input size of 224. The dimension of the extracted representational embeddings was 512. Furthermore, we replaced the original video transformer with the proposed Visual Fusion Module. For the audio baseline, we adopted a wav2vec 2.0 (Baevski et al., 2020b) model that was pre-trained with 960 hours of Librispeech on 16kHz sampled speech audio within the SEGUE pre-training method for SLU. Since SLU is a different downstream task than ours, we used the output of the nine transformer layers in the contextualized encoder of wav2vec 2.0.

4.2 Video emotion recognition performance

We evaluated the proposed visual fusion module (VFM) and the fine-tuning of emotion prompts in two steps. First, we used the same linear-probe evaluation method adopted in CLIP (Radford et al., 2021) to directly show the quality of the fine-tuned emotion representations. The proposed VFM helped the original CLIP structure to achieve 50.37% accuracy and 37.13% weighted F1 score. Then, we further evaluated the supervised performance. In particular, we compared the performance of our fine-tuned model followed by a few-shot trained classifier (50.37% and 37.13%) and other supervised ERC methods. In addition, we added an ablation experiment within each step to show the effect of prompt can further improve 1.01% accuracy and 0.82% weighted F1 score.

The linear-probe evaluation results of video-based ERC are shown in Table 2: our methods achieved better performance than other CLIP-based methods in terms of accuracy and weighted F1 score. Our first experiment (VFM) trained VFM only with utterance-only text guidance. The VFM+prompt trained VFM with

TABLE 3 The performance of the supervised fine-tuned visual fusion module (VFM) and classifier compared with other existing supervised methods in terms of accuracy and weighted F1 score.

Method	Supervised	
	Acc	Weighted F1
GRAPHCFC (Li et al., 2024)	47.59	33.26
EmoCaps (Li et al., 2022)	31.64	31.26
M2FNet (Chudasama et al., 2022)	45.63	32.44
VFM	50.83	38.28
VFM + prompt	**52.04** + 1.21	**41.18** + 2.90

The bold value is the highest value.

TABLE 4 The performance of the supervised fine-tuned auditory adapter (AA) and classifier compared with other existing supervised methods in terms of accuracy and weighted F1 score.

Method	Supervised	
	Acc	Weighted F1
CFN-ESA (Li et al., 2023)	49.35	41.46
GRAPHCFC (Li et al., 2024)	47.55	**41.62**
M2FNet (Chudasama et al., 2022)	49.04	39.63
AA	51.72	39.08
AA + prompt	**52.66** + 0.94	41.50 + 2.42

The bold value is the highest value.

the emotion prompt proposed in Section 3.2.2. The results show that the emotion prompt can further improve the performance of VFM. Compared with other video-level CLIP methods, the main difference of our proposed VFM is that we adopted the transformer decoder structure and trainable embeddings to extract emotion-related features instead of only considering the temporal or contextual information.

The evaluation results for the supervised methods are shown in Table 3. We adopted the few-shot emotion classifier proposed in Section 3.4 to categorize the extracted visual emotion representations instead of the linear classifier. The emotion classifier uses the extracted representations to construct its cache model. When the classifier is used to improve the VFM method, it will use the representations extracted from VFM, achieving 50.83% accuracy and 38.28% weighted F1 score. When it works for VFM+prompt, the cache model of the classifier will be updated by the representations extracted from VFM+prompt and achieve 52.04% accuracy and 41.18% weighted F1 score. The results show that our methods can outperform the existing supervised multi-modal methods when only working on video data. Compared with Table 2, the emotion classifier can further improve the emotion recognition accuracy of both VFM and VFM+prompt methods by achieving 1.21% accuracy and 2.90% weighted F1 score.

4.3 Emotion recognition performance in audio and two modalities

We evaluated our transferred wav2vec 2.0 on MELD to show the audio-based emotion recognition performance. The cross-modality pre-training study on wav2vec 2.0 did not receive as much attention as the CLIP-based visual pre-trained models. Therefore, we only compared our audio-based method with the existing supervised method. The results in Table 4 report that the proposed methods can outperform the existing supervised methods in terms of accuracy (52.66%) and can achieve comparable weighted F1 scores (41.50%) to these supervised methods. The ablation comparison shows that both the transfer learning auditory adapter and the prompt fine-tuning method can improve the emotion recognition performance of audio representations.

In addition, we also conducted an evaluation of the joint emotion classifier on the MELD dataset to assess its performance on the joint visual and acoustic representations in Table 5. The results presented in Table 5 demonstrate that our proposed methods have superior accuracy (51.86 + 2.29%) compared to existing supervised two-modalities methods. Additionally, our methods achieved weighted F1 scores (42.34 + 0.46%) that surpass those of the majority of other methods, with the exception of CFN-ESA (Li et al., 2023) (the best weighted F1 score is 43.25%).

It is worth noting that the multi-modal input did not provide any further improvement over the single-modal baseline in Tables 3, 4. On the contrary, the accuracy was relatively reduced compared to the video and audio-based baselines. We believe that the unsatisfactory multi-modal performance was caused by the following reasons: (1) *Indirect feature alignment.* The visual and acoustic representations were only trained with align to the same text representations but were not aligned with each other. This indirect alignment led to insufficient synchronization of the representations from different modalities. (2) *No trainable multi-modal fusion module.* The adopted joint cache model was not a trainable feature fusion module but only a decision ensemble module with two hyper-parameters to balance the different modalities. However, it is essential to make a trade-off between the performance of the experimental dataset and the practical application scenario. Although a trainable late fusion classification module could improve the testing performance on the MELD dataset, it could also face challenges related to distribution shifts in real-world application scenarios. In contrast, the few-shot cache model-based classifier offered greater adaptability as it could be readily updated by collecting a limited K-shot training dataset from the specific target application scenarios.

5 Future work

The proposed methods adapt the pre-trained model to a specific emotion recognition scenario—AI-powered assistive technology for children with ASD. In this work, we focused on designing the core deep learning-based emotion recognition model for the whole assistive technology system. Our emotion recognition assistive

TABLE 5 The performance of the supervised fine-tuned two modalities fusion and joint classifier compared with other existing supervised methods in terms of accuracy and weighted F1 score.

Method	Supervised	
	Acc	Weighted F1
CFN-ESA (Li et al., 2023)	50.34	**43.25**
GRAPHCFC (Li et al., 2024)	47.61	41.62
M2FNet (Chudasama et al., 2022)	48.35	35.74
VFM + AA	49.57	41.88
VFM + AA + prompt	**51.86** + 2.29	42.34 + 0.46

The bold value is the highest value.

technology aims to provide daily life assistance for helping autistic children understand others' emotions in conversation.

Toward this goal, a novelty presented in this work was the use of a k-shot cache model as an emotion recognition classifier. The benefit of the cache model is to accelerate the fine-tuning speed, which can efficiently improve the model's real-world application performance. In a future system, we will design an interface for the assistance system so that the parents or supervisors of autistic children can quickly adapt the model to children's appropriate living environments and familiar conversational partners. For example, the system will automatically record a few video clips of conversations that are ambiguous for emotion recognition. Then, parents can update the cache model by providing the correct labels and fine-tuning. In this case, the emotion recognition assistive function can be improved and customized for the users' living environment.

6 Conclusions

In this paper, we have presented a transfer learning framework, EmoAsst, to exploit the power of large pre-trained models, specifically CLIP and wav2vec 2.0, to significantly advance video and audio-based ERC tasks. Through our EmoAsst, we have effectively increased the performance of emotion recognition in both audio and video domains, guided by the knowledge from text-based methods. The EmoAsst includes the development of a visual fusion module and an emotion prompt fine-tuning method, both of which have successfully enriched CLIP's visual emotion representations by incorporating text-based guidance. Additionally, we have demonstrated the efficacy of adopting CLIP's text encoder and applying supervised contrastive learning to enhance the transfer learning process for wav2vec 2.0, resulting in improved audio-based ERC. To further improve the accuracy of video and audio-based ERC, we have introduced a novel joint few-shot emotion classifier that takes advantage of the fine-tuned visual and acoustic representations. The evaluation of the MELD dataset underscores the exceptional performance of our methods, which outperform the majority of existing video and audio-based approaches. Importantly, our work has shown that all of the proposed methods have the potential to improve video and audio-based ERC.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://affective-meld.github.io/.

Author contributions

MW: Methodology, Writing – original draft, Data curation, Formal analysis, Investigation, Resources, Software, Validation, Visualization. NY: Methodology, Writing – original draft, Funding acquisition, Project administration, Supervision, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work was supported in part by NY's startup funding.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Baevski, A., Zhou, H., Mohamed, A., and Auli, M. (2020a). "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS'20* (Red Hook, NY: Curran Associates Inc).

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020b). wav2vec 2.0: a framework for self-supervised learning of speech representations. *Adv. Neural Inform. Proc. Syst.* 33, 12449–12460.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., et al. (2008). Iemocap: Interactive emotional dyadic motion capture database. *Lang. Res. Evaluat.* 42, 335–359. doi: 10.1007/s10579-008-9076-6

Chen, S.-Y., Hsu, C.-C., Kuo, C.-C., and Ku, L.-W. (2018). Emotionlines: an emotion corpus of multi-party conversations. *arXiv [Preprint]*. arXiv:1802.08379v2. doi: 10.48550/arXiv.1802.08379

Chudasama, V., Kar, P., Gudmalwar, A., Shah, N., Wasnik, P., and Onoe, N. (2022). "M2fnet: Multi-modal fusion network for emotion recognition in conversation," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (New Orleans, LA: IEEE), 4651–4660.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An image is worth 16x16 words: transformers for image recognition at scale. *arXiv [Preprint]*. arXiv:2010.11929v2. doi: 10.48550/arXiv.2010.11929

Fan, Y., Lu, X., Li, D., and Liu, Y. (2016). "Video-based emotion recognition using CNN-RNN and c3d hybrid networks," in *The Proceedings of the 18th ACM International Conference on Multimodal Interaction* (New York, NY: ACM), 445–450. doi: 10.1145/2993148.2997632

Febrian, R., Halim, B. M., Christina, M., Ramdhan, D., and Chowanda, A. (2023). "Facial expression recognition using bidirectional LSTM - CNN," in *Procedia Computer Science, 216:39-47. 7th International Conference on Computer Science and Computational Intelligence 2022.*

Feng, K., and Chaspari, T. (2020). A review of generalizable transfer learning in automatic emotion recognition. *Front. Comp. Sci.* 2, 9. doi: 10.3389/fcomp.2020.00009

Ghaleb, E., Popa, M., and Asteriadis, S. (2019). "Multimodal and temporal perception of audio-visual cues for emotion recognition," in *Proceedings of the 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)* (Cambridge: IEEE), 552–558. Available online at: https://api.semanticscholar.org/CorpusID:207993833

John, V., and Kawanishi, Y. (2022). "Audio and video-based emotion recognition using multimodal transformers," in *2022 26th International Conference on Pattern Recognition (ICPR)*, 2582–2588.

Kahou, S. E., Michalski, V., Konda, K., Memisevic, R., and Pal, C. (2015). "Recurrent neural networks for emotion recognition in video," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction* (New York, NY: ACM), 467–474. doi: 10.1145/2818346.2830596

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., et al. (2020). Supervised contrastive learning. *Adv. Neural Inf. Process. Syst.* 33, 18661–18673. doi: 10.48550/arXiv.2004.11362

Li, J., Liu, Y., Wang, X., and Zeng, Z. (2023). CFN-ESA: a cross-modal fusion network with emotion-shift awareness for dialogue emotion recognition. *arXiv [Preprint]*. arXiv:2307.15432v1. doi: 10.48550/arXiv.2307.15432

Li, J., Wang, X., Lv, G., and Zeng, Z. (2024). Graphcfc: a directed graph based cross-modal feature complementation approach for multimodal conversational emotion recognition. *IEEE Trans. Multimedia.* 77–89. doi: 10.1109/TMM.2023.3260635

Li, Z., Tang, F., Zhao, M., and Zhu, Y. (2022). Emocaps: emotion capsule based model for conversational emotion recognition. *arXiv [Preprint]*. arXiv:2203.13504v1. doi: 10.48550/arXiv.2203.13504

Lin, Z., Geng, S., Zhang, R., Gao, P., de Melo, G., Wang, X., et al. (2022). "Frozen clip models are efficient video learners," in *European Conference on Computer Vision* (Cham: Springer), 388–404.

Losh, M., and Capps, L. (2006). Understanding of emotional experience in autism: Insights from the personal accounts of high-functioning children with autism. *Dev. Psychol.* 42, 809–818. doi: 10.1037/0012-1649.42.5.809

Loshchilov, I., and Hutter, F. (2017). "Decoupled weight decay regularization," in *arXiv*.

Ma, F., Zhang, W., Li, Y., Huang, S.-L., and Zhang, L. (2020). Learning better representations for audio-visual emotion recognition with common information. *Applied Sci.* 10, 20. doi: 10.3390/app10207239

Ma, Y., Xu, G., Sun, X., Yan, M., Zhang, J., and Ji, R. (2022). "X-clip: end-to-end multi-grained contrastive learning for video-text retrieval," in *Proceedings of the 30th ACM International Conference on Multimedia* (Lisbon: ACM), 638–647.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). Pytorch: an imperative style, high-performance deep learning library. *Adv. Neural Inform. Process. Syst.* 8024–8035. doi: 10.48550/arXiv.1912.01703

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

Radford, A., Kim, J., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., et al. (2021). Learning transferable visual models from natural language supervision. *arXiv [Preprint]*. arXiv:2103.00020v1. doi: 10.48550/arXiv.2103.00020

Rasheed, H., Khattak, M. U., Maaz, M., Khan, S., and Khan, F. S. (2023). "Fine-tuned clip models are efficient video learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Vancouver, BC: IEEE), 6545-6554.

Tager-Flusberg, H. B. (2000). Understanding the language and communicative impairments in autism. *Int. Rev. Res. Ment. Retard.* 23, 185–205. doi: 10.1016/S0074-7750(00)80011-7

Tan, Y. X., Majumder, N., and Poria, S. (2023). "Sentence Embedder Guided Utterance Encoder (SEGUE) for spoken language understanding," in *Proc. INTERSPEECH* (Dublin: ISCA), 3914–3918. doi: 10.21437/Interspeech.2023-1392

Tashu, T. M., Hajiyeva, S., and Horvath, T. (2021). Multimodal emotion recognition from art using sequential co-attention. *J. Imag.* 7, 8. doi: 10.3390/jimaging7080157

Xu, H., Ghosh, G., Huang, P.-Y., Okhonko, D., Aghajanyan, A., Metze, F., et al. (2021). Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv [Preprint]*. arXiv:2109.14084v2. doi: 10.48550/arXiv.2109.14084

Zhang, R., Fang, R., Zhang, W., Gao, P., Li, K., Dai, J., et al. (2021). Tip-adapter: training-free clip-adapter for better vision-language modeling. *arXiv [Preprint]*. arXiv:2111.03930v2. doi: 10.48550/arXiv.2111.03930

Zhang, S., Pan, Y., and Wang, J. Z. (2023). "Learning emotion representations from verbal and nonverbal communication," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Vancouver, BC: IEEE), 18993–19004.

Zhang, S., Zhang, S., Huang, T., Gao, W., and Tian, Q. (2018). Learning affective features with a hybrid deep model for audio-visual emotion recognition. *IEEE Trans. Circuits Syst. Video Technol.* 28, 3030–3043. doi: 10.1109/TCSVT.2017.2719043