CLMER: a Framework for Contrastive Learning-based Multi-modal Emotion Recognition

Anonymous

***Abstract*—Emotion recognition plays a crucial role in human-computer interaction and affective computing, yet its effectiveness is limited by the difficulty of integrating heterogeneous modalities with fundamentally different structures, such as physiological signals and visual data. In this paper, we propose CLMER, a contrastive learning-based multi-modal cross-attention framework designed to address the challenges of complex emotion recognition. The framework introduces a serialization strategy that converts pixel-level image data into time-series data, aligning it with the temporal characteristics of physiological signals. CLMER consists of three core components that work together to enable effective multi-modal emotion recognition. The multi-modal data preparation module preprocesses physiological and visual data, ensuring consistency across modalities. Building on this foundation, the contrastive learning-based feature extraction module generates temporal representations that capture the essential patterns embedded in the data through self-supervised learning. Finally, the multi-modal fusion module employs cross-modal attention to integrate features with improved modality alignment. Experimental evaluations on two public datasets DEAP, AMIGOS and a private dataset MAN-II demonstrate that CLMER significantly outperforms unimodal and traditional fusion approaches, achieving state-of-the-art performance in emotion classification tasks. These findings highlight the framework's robust generalization, computational efficiency, and strong performance in multi-modal emotion recognition, suggesting its potential for real-world deployment. Our code is available at https://github.com/anonymous/CLMER.**

***Index Terms*—multi-modal emotion recognition, contrastive learning, data serialization, cross-modal attention.**

I. Introduction

HUMAN emotion recognition serves as a foundational task in human-computer interaction and affective brain-computer interfaces, enabling the personalized and adaptive user experiences and supporting applications in domains such as healthcare and gaming [1]. With the increasing demand in these fields, various deep learning methods have been introduced to provide more efficient and generalized support for emotion recognition technologies [2]. Emotion recognition can be conducted using various types of materials, among which are behavioral data such as facial expressions [3], body posture [4], and vocal attributes [5] to train deep learning models. However, subjects may intentionally or unintentionally distort these cues by displaying misleading expressions or modulating their voice [6]. A more reliable and less easily fabricated approach relies on physiological signals for emotion recognition. For instance, data such as Electroencephalography (EEG) [7], Electrocardiography (ECG) [8], and Galvanic Skin Response (GSR) [9], collected via wearable devices, are commonly used in deep learning-based classification models.

Given the complex and dynamic nature of emotions, unimodal approaches often suffer from inherent limitations and narrow perspective on emotion, reducing model robustness and generalization capacity. Recent studies have increasingly focused on the exposition of multi-modal approaches, with various fusion methods demonstrating promising improvements in training performance [10]. Although multi-modal data provide a more comprehensive representation of human emotional states, prior research has largely concentrated on either internal expression, such as EEG and ECG [11], or external manifestations such as audio and images [12]. However, approaches that combine both internal and external modalities, for instance, physiological signals with visual data, remain limited and have yet to employ state-of-the-art deep learning models [1]. We formulate a strategy that fuses visual information with EEG-centered physiological signals for multi-modal emotion recognition based on the mechanism of cross-modal attention and contrastive learning.

Heterogeneous modalities embody diverse qualities, structures, and representations, which pose significant challenges for both learning modal-specific representations and capturing alignment during the fusion process [13]. In terms of representation learning, contrastive learning methods effectively extract key feature information from each modality based on task requirements, enabling further training and fusion of the model. A contrastive learning-based fusion method, termed CILP, jointly trains the image encoder and the text encoder for prediction that shows outstanding performance while being computationally efficient [14]. In addition, related studies have shown that incorporating self-supervised learning methods enhances training performance in the fusion of speech and text modalities [15]. To better address the significant structural differences between the physiological signal modality and the visual modality, we adopted a deep fusion strategy that uses contrastive learning methods to extract features and facilitate the fusion process.

To effectively align data from both modalities for fusion, we propose a serialization strategy to the visual data in the temporal dimension. This design is motivated by the temporal nature of physiological signals. Unlike traditional approaches that use pixel-based images as input to deep learning models [16], we extract facial information, such as 3D landmarks [17] and Action Units (AU) [18] from images as visual input. This allows the original images to be reformatted and subsequently aligned with the physiological signal modality along the temporal dimension. The serialization strategy that we proposed not only mitigates the various impacts caused by structural differences between the modalities, but also eliminates the need to separately address pixel-based image and sequence structures when designing feature extraction. By focusing solely on the temporal structure, we design

a contrastive learning method tailored for time-series data, applicable to both modalities.

In this paper, we propose a novel framework for **C**ontrastive **L**earning-based **M**ulti-modal **E**motion **R**ecognition (CLMER), which can effectively integrate the internal modality represented by EEG-based physiological signals and the external modality represented by visual data. The framework demonstrates notable robustness and generalization capabilities. The framework is primarily made up of three components arranged in a sequential linear process. The first component is dedicated to the preprocessing of the two modalities. Specifically, standard signal processing operations are performed on EEG-dominant physiological signals to provide the framework with reliable and high-quality physiological data. Serialization is applied to visual data to align them with the physiological signal modality along the temporal dimension. In the middle stage of the framework, after applying data augmentation to both modalities, contrastive learning is employed to further extract temporal features and inter-sample characteristics from two modalities. Taking advantage of the serialization design of the visual modality, the contrastive learning method only needs to focus on the sequential structure. In the final stage, the features of the two modalities are fused to complete the emotion recognition and classification task. The primary contributions of this work can be summarized as follows:

1) We propose a novel contrastive learning-based multi-modal fusion framework for emotion recognition. By integrating internal modalities (e.g., physiological signals) with external modalities (e.g., visual data), the framework achieves robust performance, strong generalization, and offers promising opportunities for further development.
2) The serialization strategy for visual data, enhancing its alignment with the temporal characteristics of physiological signals. This enables visual data to better support the temporally specialized Contrastive Learning method while also streamlining the overall framework design, leading to reduced data volume and improved training efficiency.
3) A contrastive learning-based feature extraction method specifically designed for time-series data enables the extraction of features from sequential multi-modal data, thereby facilitating deeper and more effective fusion.

After presenting relevant research and background information on multi-modal fusion for emotion recognition (see Section II-A) and contrastive learning (see Section II-B), we propose the CLMER framework. We provide a detailed explanation of the specific mechanisms and processes of data serialization (see Section III-B2), contrastive learning-based feature extraction (see Section III-C) and multi-modal fusion (see Section III-D) within the framework. The experimental evaluation is conducted on the publicly available AMIGOS and DEAP datasets, as well as our proprietary MAN2 dataset (see Section IV-A). In the proposed framework, all datasets show strong performance, achieving notable improvements in fusion methods (see Section IV-D & IV-E). A detailed discussion of the parameters within the framework was also conducted (see Section IV-F).

II. Related Works

A. Emotion Recognition

Emotion recognition is typically accomplished by addressing classification tasks using deep learning methods. These classification tasks are based on various emotion paradigms, which are primarily categorized into two types: discrete and multidimensional [1]. The discrete paradigm, exemplified by Ekman's basic emotion theory, includes six fundamental emotions: happiness, sadness, anger, fear, surprise, and disgust [19] [20], whereas the multidimensional paradigm is represented by Russell's theory of a two-dimensional continuous space characterized by valence and arousal [21].

There has been relatively extensive studies on unimodal emotion recognition, with four primary modalities being widely used: speech signals [22], text [23], facial expressions, and physiological signals [24]. In the domain of physiological signals [25], Zhang et al. applied an attention mechanism to both channel and temporal dimensions of EEG data, achieving strong performance on both public and private datasets [26]. In terms of ECG, Fan et al. integrated the channel and spatial dimensions to explore multi-dimensional fusion, thereby enhancing emotion recognition accuracy [8]. For visual modality, Jain et al. used a CNN to extract features from raw facial images, followed by a Recurrent Neural Network (RNN) to spread information and perform emotion classification [27]. It can be observed that recent studies increasingly tend to take advantage of advanced attention mechanisms to better model the sequential nature of physiological signals.

Human emotional states are complex and dynamic, making it challenging for single-modal data to capture them both reliably and accurately. This issue is particularly prominent in external modalities, as such data can be influenced or misled by subjects through deliberate behaviors [6]. Therefore, multi-modal fusion approaches are increasingly adopted to integrate data from multiple modalities and diverse perspectives, enhancing the diversity and volume of data and further improving the stability and authenticity of emotion recognition [28]. Tsai et al. employed a cross-modal attention mechanism during the fusion process to address the issue of non-alignment among vision, text, and audio modalities [29]. Another study adopted a similar mechanism to mainly fuse various structurally similar physiological signals, including EEG and ECG, achieving favorable results [30]. Only a limited number of studies have explored the fusion of EEG and visual data. Earlier work by Koelstra et al. extracted features from both modalities and then compared the effectiveness of feature-level fusion and decision-level fusion [31]. Recently, Hosseini et al. combined quantitative and qualitative modes to perform binary classification [32] for each dimension of the 3D emotional space model [33]. Overall, the fusion of EEG and visual data in emotion recognition remains underexplored, with most existing models based on CNNs, RNNs, or LSTMs [1]. Moreover, the use of attention mechanisms for this fusion is rare, potentially limiting the efficient exploitation of temporal and sequential information present in both modalities for emotion recognition tasks.

B. Contrastive Learning

Self-supervised contrastive learning methods have been widely applied in the field of image analysis. For example, Chen et al. proposed A **Sim**ple Framework for **C**ontrastive **L**earning of Visual **R**epresentations (SimCLR), which calculates the loss of positive pairs based on the concept of maximization of mutual information (MI) [34]. Building on SimCLR, Mohsenvand et al. further adapted a similar approach to sequential data, introducing **Seq**uential **C**ontrastive **L**earning of **R**epresentations (SeqCLR). They applied channel augmentation to EEG signals and calculated the normalized temperature-scaled cross-entropy (NT-Xent) loss. Further self-supervised methods for temporal data include that of Liu et al., who mixed two augmentations of samples at the same timestamp to form double Universums, which were then used to construct a loss function [35]. Eldele et al. applied two types of augmentations to the temporal data of the EEG: strong augmentation that introduces significant perturbations while retaining partial temporal information, and weak augmentation that applies minor changes to the original signal. They used a contrastive learning loss function designed to predict future information based on past time-series data [36].

The aforementioned contrastive learning methods for time-series data have achieved promising in EEG signal analysis. We consider extending these approaches to other modalities with similar structures. In particular, we serialize the image modality into time-series data, which aligns its structure with that of EEG signals. This transformation not only reduces the overall data size, thereby improving training efficiency, but also enables the use of identical contrastive learning techniques across both modalities. Moreover, the feature extraction process in our framework becomes more streamlined, as time-series–oriented methods can be applied directly without additional modality-specific adjustments.

III. Methods

A. Overview

In this section, we present the specific structural composition of CLMER, our proposed framework for emotion recognition (see Figure 1). The core of the framework consists of three main modules: a visual data serialization module, a contrastive feature extraction module, and a multi-modal fusion module. Specifically, we extract facial features from visual data to convert them into a time-series format consistent with physiological signals. The sequence data from each modality are then fed into a contrastive learning model for feature extraction. In this module, we employ self-supervised contrastive learning methods tailored for time-series data extracting useful representations from unlabeled data. Both visual and signal modalities share a unified feature extraction process, ensuring a compact and simple framework design. The features obtained from contrastive learning are subsequently applied to the emotion recognition task. Finally, the multi-modal fusion module employs cross-modal attention mechanism, training modalities in pairwise and utilizing the resultant tensors for the classification task. The CLMER framework leverages the advantages of multi-modal fusion in terms of accuracy and robustness over unimodal approaches. Furthermore, by employing serialized visual data in conjunction with contrastive learning for feature extraction, it reduces the negative effects of structural discrepancies between different modalities which, in our study, are the differences between visual and physiological signals, thereby enhancing the efficacy of the fusion process.

B. Multi-modal Data Preparation Module

1) Physiological Signal Preprocessing Module: For EEG or physiological signals, we adopted a simple strategy to streamline the framework, which involves controlling the sampling frequency and applying signal filtering to preliminarily preserve the distinctive information and features of the raw data. For physiological signals as input, we first standardize the sampling frequency to a specific value, and then apply bandpass filter to remove low-frequency baseline drift and high-frequency noise.

2) Visual Data Serialization Module: Unlike previous methods in emotion recognition research that use facial images in raw pixel pattern, our approach does not treat facial images as direct inputs for the visual modality. Instead, we extract meaningful visual information from these images (converted from videos) and represent it in a time-series pattern, ensuring consistency with EEG signal data. Before employing contrastive learning methods for feature extraction, we adopt a model structure partly derived from OpenFace [37] to extract visual information in facial images and then integrate them for the purpose of serialization. We selected two representative types of facial visual data for emotion recognition.

- The first type consists of 3D landmark coordinates, comprising 68 points strategically distributed across key facial features including the eyes, mouth, and nose. The temporal dynamics of these 3D landmarks allows us to capture the trends in the movements of different facial regions with greater precision.
- The second type is Action Units (AUs), serving as supplementary explanatory data for 3D landmarks, which are components of the Facial Action Coding System (FACS) designed to classify and encode human facial movements. AUs are obtained through a detailed analysis of individual facial muscle activations, enabling the decomposition of any anatomically possible facial expression into its specific units.

We organized the extracted semantic information from the images into a temporal data structure, thereby achieving the serialization of visual data. This method aligns visual and EEG data along the temporal dimension, providing input tensors to the feature extraction module. The serialization approach can also be applied to other types of facial visual data, such as gaze tracking and HOG features. In the experiment section of our study, we select only two visual features.

C. Contrastive Learning Based Feature Extraction Module

To further enhance the fusion efficiency, accuracy, and robustness of the two modalities, we add a feature extraction module following the serialization step. We use the self-supervised **C**ontrastive **L**earning (CL) methods specifically

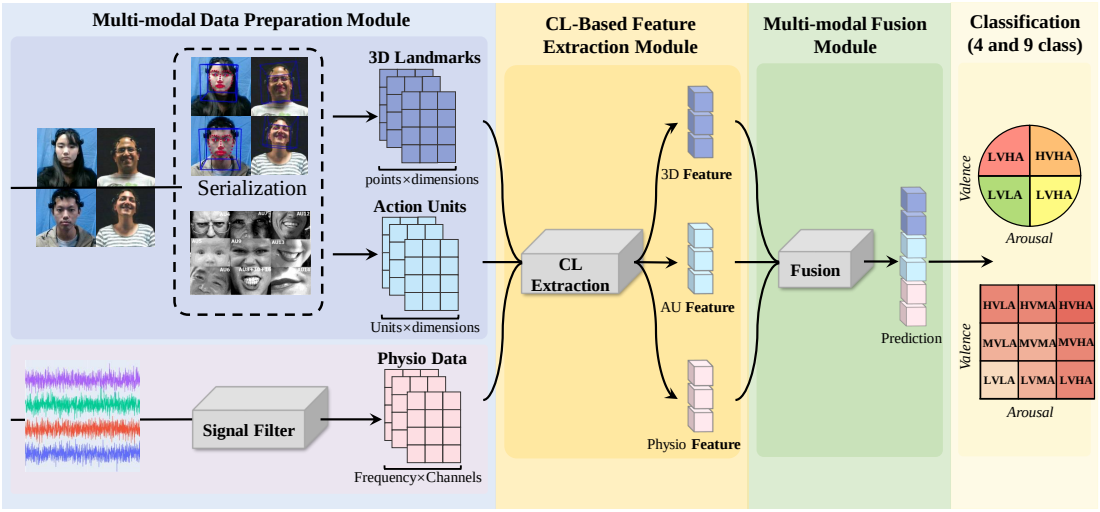


Fig. 1: The overall framework of the proposed CLMER, which is primarily composed of three components: the Multi-modal Data Preparation Module, the CL-Based Feature Extraction Module, and the Multi-modal Fusion Module.

designed for processing time-series data. In this module, sequential features in temporal dimension and latent features among samples are captured, generating discriminative representations for subsequent fusion. In addition to processing the physiological data, we also adapted these methods for the serialized visual data (see Figure 2).

Contrastive learning constructs loss functions based on the theory of MI [34]. The objective of the loss function is to maximize the similarity among augmentations originating from the same sample or a designated group of samples while minimizing the similarity between augmentations derived from different samples or distinct groups [38]. Similarity between augmentations can be measured using dot product or cosine similarity etc. The SeqCLR [39] method, which adapts the contrastive learning approach SimCLR [40] from image data to time-series data, can effectively learn representations from EEG data. In our framework, techniques specifically designed for processing electrical signals, such as scaling and jittering, are employed for data augmentation to generate augmented views. Similarity calculations are then performed. Loss functions are outlined in Eq.(1) and Eq.(2), (z^a, z^b) represents positive pair and (z^a, z^c) represent negative pairs. τ is temperature parameter, $\mathbb{1}_{[j\neq i]} \in \{0, 1\}$ is 1 iff $j \neq i$.

$$\ell(a,b) = -log\frac{exp(sim(z^a, z^b)/\tau)}{\sum_{c=1}^{2N} \mathbb{1}_{[c\neq i]} exp(sim(z^a, z^c)/\tau)} \quad (1)$$

$$\mathcal{L}_{sample} = \frac{1}{2N}\sum_{c=1}^{2N}[\ell(2c-1, 2c) + \ell(2c, 2c-1)] \quad (2)$$

Another strategy we adopted to obtain temporal representations is using the summaries s_t of samples $z_{[j,\ldots,t]}$ within a certain period of time $\mathbb{T}$ to predict sample z_{t+i} in the future timeline, as shown in Eq.(3), based on the same MI theory. The loss function aims to maximize the similarity between the prediction and the future target, thus bringing the two closer while minimizing the prediction with other samples in the given period of time to train the summarization process

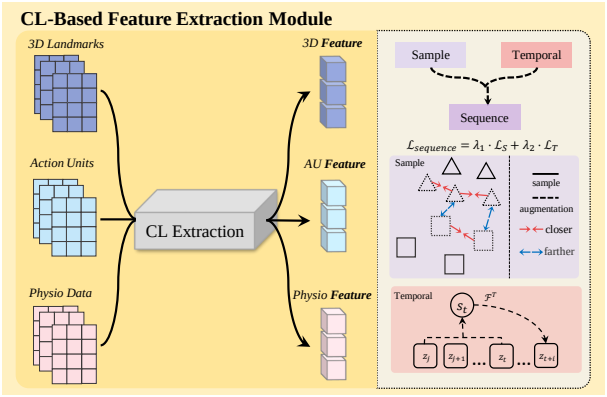


Fig. 2: The specific details of the CL-Based Feature Extraction Module and the working principle of its corresponding loss function. The Sequence Loss is composed of two parts: sample Loss and temporal Loss.

[36]. $\mathcal{F}$ is a function transform s_t into the same dimension as z aligning the prediction with samples.

$$\mathcal{L}_{temporal} = -\frac{1}{M}\sum_{i=1}^{M} log\frac{exp(sim(\mathcal{F}^T(s_t), z_{t+i}))}{\sum_{j\in\mathbb{T}\setminus\{t\}} exp(sim(\mathcal{F}^T(s_t), z_j))} \quad (3)$$

By integrating $\mathcal{L}_{sample}$ and $\mathcal{L}_{temporal}$, we can construct a more stable loss function Eq.(4) that facilitates learning more robust representations [14] in temporal dimension. λ_1 and λ_2 are loss weighting hyper-parameters.

$$\mathcal{L}_{sequence} = \lambda_1 \cdot \mathcal{L}_{sample} + \lambda_2 \cdot \mathcal{L}_{temporal} \quad (4)$$

D. Multi-modal Fusion Module

In our model, we employ a method fusing modalities in pair capable of accommodating various data modalities as inputs (see Figure 3). Therefore, this approach is well-suited for the physiological and visual fusion strategy proposed in this paper.

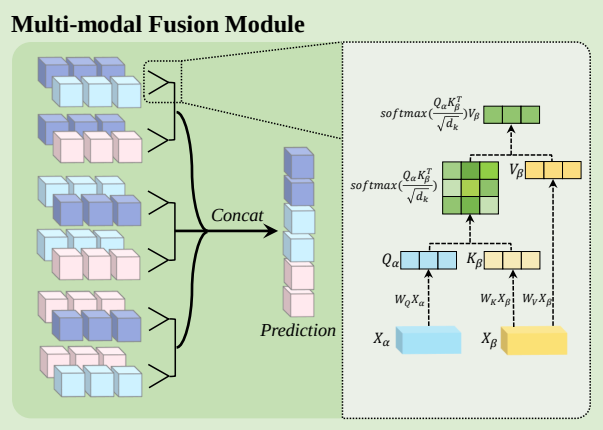


Fig. 3: The specific details of the Multi-modal Fusion Module and the operational process of the cross-modal attention mechanism for fusing data from two modalities

Considering that the data from various modalities in this study have been unified into a sequential structure, employing attention-based models for training can further improve the framework's performance. We utilize the cross-modal attention mechanism for the multi-modal fusion of different modalities, which incorporates certain improvements upon the original attention principles, making it more suitable for training with multi-modal data. Cross-modal attention is modified on the basis of original attention mechanism that establishes a connection between two modalities for computation. In the following equation, Q_α represents the $Query$ provided by modality α, which is obtained through the transformation $X_\alpha W_{Q\alpha}$. Meanwhile, $K_\beta(Key)$ and $V_\beta(Value)$ correspond to the input sequences from the other modality, calculated via $X_\beta W_{K\beta}$ and $X_\beta W_{V\beta}$, respectively. As shown in Eq.(5) and Eq.(6), the remaining components of the equation remain unchanged and are consistent with the original mechanism, referred to as the single-head cross-modal attention.

$$Y^{head_i}_{(\alpha\to\beta)} = Attention(Q_\alpha, K_\beta, V_\beta) \tag{5}$$

$$Y^{head_i}_{(\alpha\to\beta)} = softmax(\frac{Q_\alpha K_\beta^T}{\sqrt{d_k}})V_\beta \tag{6}$$

Similar to the original mechanism, as demonstrated in Eq.(7), the outputs of n single-head attentions are concatenated to obtain the output of the multi-head cross-modal attention.

$$Y^{mult}_{(\alpha\to\beta)} = Concat(Y^{head_1}_{(\alpha\to\beta)}, ..., Y^{head_n}_{(\alpha\to\beta)}) \tag{7}$$

IV. Experiments and Result

A. Datasets

1) DEAP Dataset: Human-machine interaction exhibits behavioral patterns remarkably similar to those observed in human-to-human interactions [6]. In support of this premise, the DEAP dataset was specifically generated to investigate human-machine interactions that effectively evoke emotional responses from participants. From the initial selection of 120 music videos from Last.fm[1], comprising 60 tracks filtered based on emotional labels and 60 manually selected tracks to ensure a balanced distribution across the valence-arousal spectrum [21], a total of 40 videos were ultimately retained. Their emotional impact was validated through linear regression analysis, assessing valence, arousal, and dominance on a 9-point scale [21]. This meticulous selection process ensured that the final music collection effectively elicited a range of emotional responses and eliminate researchers' subjective emotional annotations, facilitating the collection of reliable EEG data for subsequent physiological signal-based training [41]. In addition to EEG data, we employed our serialization module to extract facial features, which include the 3D coordinates of facial landmarks [17] [42] and facial action units [18] from the video recordings provided by the DEAP dataset. To categorize emotional responses more effectively, we divided the valence and arousal indicators into three distinct scales: low [1,3), medium [3,6), and high [6,9]. By combining these six scales in pairs, we generated nine classification labels that segment emotions in a more precise way including LVLA (low valence low arousal), LVMA (low valence medium arousal), LVHA (low valence high arousal), MVLA (medium valence low arousal), MVMA (medium valence medium arousal), MVHA (medium valence high arousal), HVLA (high valence low arousal), HVMA (high valence medium arousal), HVHA (high valence high arousal).

2) AMIGOS Dataset: The AMIGOS dataset includes a collection of 16 short emotional videos, each designed to elicit specific affective states in individual viewers. These videos, ranging from 51 to 150 seconds, were carefully selected based on their valence and arousal ratings, as annotated by 72 participants. The categorization of these videos into quadrants of the valence-arousal space allows for a nuanced examination of emotional responses [21] [43]. The experimental design ensures that each video is shown in a controlled environment, enhancing the reliability of the recorded neuro-physiological signals, such as EEG, ECG, and GSR. This multi-modal approach aligns with DEAP that emphasize the importance of using diverse stimuli to capture affective responses effectively [41]. The short video segment of the AMIGOS dataset serves as a vital resource for our experiment, we use EEG and facial features to train CLMER. We classified each of the videos into one of four quadrants of the valence-arousal (VA) space align with AMIGOS's classification [44] including LVLA (low valence low arousal), LVHA (low valence high arousal), HVLA (high valence low arousal), HVHA (high valence high arousal).

In addition to the short video data, the AMIGOS dataset features four long videos, each exceeding 14 minutes in duration. The long video component of the AMIGOS dataset encompasses a comprehensive study involving 37 participants, with a particular emphasis on the individual data collected during this experiment. While three participants (IDs 8, 24, and 28) were unavailable, 17 individuals completed the experiment in isolation. Following the long video experiment, participants

[1]www.last.fm which is a platform where users assign specific emotional tags to songs.

were prompted to complete online questionnaires assessing Personality Traits [45] and the Positive and Negative Affect Schedule (PANAS) [46]. This dataset serves as a robust resource for analyzing individual emotional responses within a controlled video viewing context. In our experiment, we selected 20 segments from long videos, extracting 5 seconds of data from each segment, which includes both EEG and facial features. Additionally, the materials are categorized into four distinct classes: LVLA, LVHA, HVLA, and HVHA. This structured approach enhances the investigation of emotional dynamics in response to varying stimuli.

3) Self-developed MAN-II Dataset: In this dataset, we utilized three images and a video clip as stimulus materials to evoke four distinct emotional states: moved, angry, nervous, and reproachful. EEG data were gathered from 19 out of 23 participants using the Emotiv EPOC wireless headset, a well-established and commercially available wearable EEG device. This headset operates at a sampling rate of 128 Hz and features a total of 14 channels, adhering to the International 10-20 system. To enhance the methodology, we developed MAN-II, which builds upon the original MAN [26] setup by introducing a new class of samples and incorporating an additional modality dimension to improve emotion detection performance. Unlike the original approach, we extracted facial features including facial landmarks and subjects' eye gaze from recorded facial videos captured during the trials. The resulting dataset includes EEG data, 3D coordinates of facial landmarks [17] [42], and 2D coordinates of eye gaze [47], all collected over a duration of 12 seconds for each emotional class. All experiments involving the MAN-II dataset were approved by the Science and Technology Ethics Committee at the authors' university.

B. Evaluation and Metrics

For each dataset, we employed a random shuffling procedure to generate five distinct datasets with varying orders, which were subsequently used to train the classification model. Specifically, regarding our proposed private dataset MAN-II, we included a total of four categories: moved, angry, nervous, and reproachful. In the case of the DEAP dataset, we categorized the data into nine distinct classes by partitioning Valence and Arousal into three separate regions and combining them pairwise to form the nine categories. Additionally, the AMIGOS Short Videos and AMIGOS Long Videos datasets comprised a total of four categories. Following the training of our proposed model, we calculated the accuracy and the F1 score for the classification task, employing their respective calculation formulas, which are detailed below.

$$Accuracy = \frac{\sum_{i=1}^{C} accuracy_i}{C} \tag{8}$$

$$accuracy_i = \frac{TP_i + TN_i}{samples_i} \tag{9}$$

In this context, $accuracy_i$ refers to the accuracy of class *i*. Additionally, *C* represents the total number of categories in the classification task, *TP* denotes the number of true positive samples that have been correctly predicted as positive, *TN* indicates the number of true negative samples that are accurately classified as negative, and *samples* refer to the total number of samples within each category.

$$F1 = 2 \times \frac{Percision \times Recall}{Percision + Recall} \tag{10}$$

$$Percision = \frac{TP}{TP + FP} \tag{11}$$

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

$$F1_{weighted} = \frac{\sum_{i=1}^{C} (F_i \times samples_i)}{\sum_{i=1}^{C} samples_i} \tag{13}$$

In the formulas for Precision and Recall, *FP* represents the number of samples that are actually negative but are incorrectly predicted as positive (false positives), while *FN* denotes the number of samples that are actually positive but are incorrectly predicted as negative (false negatives). Due to the unequal distribution of sample sizes within each category across all selected datasets based on our classification method, the weighted F1 score is employed to more accurately reflect the performance of the classification task.

C. Implementation Details

Visual features, including 3D landmarks and Action Units, were extracted from each dataset after processed by our serialization module. These visual features were then combined with the EEG data to form the dataset structure used for training and evaluating the model. The proportions of the training set, validation set, and test set were approximately 70%, 15%, and 15%, respectively. For each dataset, the experiments were repeated five times; while the seed values for model training remained constant, the order of the data was randomized. The model performance was quantified by presenting both the accuracy and the weighted F1 score on the test set, along with their respective mean and standard deviation. In physiological signal preprocessing module, we first set the sampling frequency of all three datasets to a consistent 128 Hz. Subsequently, we apply a bandpass filter ranging from 4 to 45 Hz for processing the DEAP and AMIGOS datasets, while a bandpass filter from 1 to 50 Hz is used for the MAN-II dataset to preliminarily remove other interference. Additionally, for the MAN-II dataset, the filter also effectively eliminated the dominant power line interference occurring at 50 Hz in China. In the data feature extraction phase based on contrastive learning, a batch size of 64 was employed across all datasets, with a total of 500 training epochs. In the subsequent phase, during the classification task using a pairwise approach, a batch size of 64 was utilized for all datasets. The optimal number of training epochs varied according to the complexity of each dataset. Specifically, the DEAP dataset required 300 training epochs, the AMIGOS short videos' dataset necessitated 300 epochs, the AMIGOS long videos' dataset required 200 epochs, and the MAN-II dataset necessitated 300 epochs.

Both stages of the model employed the Adam optimizer. For the contrastive learning model, the learning rate was set

at 3e-4 with a weight decay of 3e-4. In contrast, the pairwise model used for the classification task had a learning rate of 1e-3 and reduced the learning rate every 20 epochs when hitting the plateau. The training process used an NVIDIA GeForce RTX 4090 GPU and PyTorch version 1.8.1 with CUDA 11.1.

D. Comparison Experiments

We selected the DEAP dataset as the primary dataset for our experiments due to its diverse emotional labels, rich and balanced emotion-eliciting materials, and comprehensive multi-modal data in large volume. As a reliable public dataset, it provides sufficient data support for our study and serves as an effective benchmark to evaluate the performance of the fusion and feature extraction modules in our framework. To address the potential complexities of real-world emotion recognition requirements, we expanded the classification standard in previous studies. Specifically, beyond performing simple binary classification on each emotional dimension, we also increased the complexity by advancing from four-class classification that combining two dimensions to nine, thereby raising the difficulty of downstream tasks for the framework.

The proposed CLMER framework primarily comprises visual data serialization, contrastive learning-based feature extraction, and multi-modal fusion modules. The methods collected for comparison with ours each have distinct characteristics. Most of the methods rely on traditional CNN or LSTM architectures for fusion. The approaches of Jung et al. employ a strategy of feature extraction followed by fusion for EEG signals and facial expressions, while HFCNN, on the other hand, focuses specifically on the fusion of EEG and peripheral physiological signal (PPS) data modalities. In standard classification tasks, CLMER achieved an accuracy and F1 score of 96.45% in the binary classification task and 96.47% in the four-class classification task, surpassing the results of other comparative methods and demonstrating exceptional performance. Despite handling more complex nine-class task compared to previous multi-modal fusion studies on the DEAP dataset, CLMER demonstrates outstanding performance among those methods reaching 96.09% in accuracy and 96.09% in F1 score (see Table I).

Results of Y.-C. Wu et al. [48], Huang et al. [49], Zhao et al. [50] and HFCNN [51] are calculated as the average of the values derived from the valence and arousal dimensions in two classes and Siddharth et al. [52] adds another liking dimension. Result of Y. Wu et al. [48] is calculated as the average of the values derived from the valence, arousal dimensions in nine classes. The paper of Hosseini et al. [32] provides two strategies for comparison. Either of them is calculated on the basis of accuracy and mean square error from valence and arousal. The rest of the results in the table refer to the original figures.

For the AMIGOS dataset, most studies have utilized the short-video mentioned in this paper (see section IV-A2), primarily conducting binary classification tasks on valence and arousal. Therefore, we applied CLMER to the same binary classification tasks on the short video dataset. Additionally, to increase task complexity, we introduced a four-class classification task, which is commonly used in other studies, to further evaluate the performance of our framework. Kamran et al. integrated patched facial data and physiological signals using an attention-based encoder for fusion. In contrast, Siddharth et al. employed a pre-trained VGG-16 model [57] to extract features from EEG transformed images, and subsequently fused the extracted features using an LSTM network. CLMER achieved an accuracy of 88.37% and an F1 score of 88.37% in the valence binary classification task on short-video data, while attaining an accuracy of 83.56% and an F1 score of 83.53% in the arousal binary classification task, surpassing the results of other comparative experiments. Furthermore, in the valence-arousal four-class classification task, CLMER obtained a notable accuracy of 81.42% and an F1 score of 81.41% (see Table II).

After expanding the original three-class emotion dataset of MAN to four-class MAN-II, our multi-modal fusion approach demonstrated improved accuracy and F1 scores compared to the single-modal FetchEEG method, even with the increased task complexity. We incorporated two attention-based multimodal fusion methods, MulT and Husformer, as comparative approaches for training on the MAN-II dataset.

TABLE I: Performance of different models on DEAP in terms of multi-class average accuracy (*Acc%*) and multi-class average F1 score (*F1%*) with stand deviations. The best results for every classification task are highlighted in **bold**.

Method	Modality	Class	Result	
			Acc	*F1*
Y.-C. Wu et al. [48]	EEG, Vision	2	68.75	–
Huang et al. [49]	EEG, Vision	2	77.27±11.29	–
Siddharth et al. [52]	EEG, Vision	2	79.60	70.00
HFCNN [51]	EEG, PPS	2	84.71	–
Zhao et al. [50]	EEG, Vision	2	86.50	–
Hosseini et al. [32]	PPS, Vision	2	96.08±0.35	–
Wang et al. [53]	EEG, Vision	2	96.89	–
Hosseini et al. [32]	EEG, Vision	2	97.59±0.22	–
CLMER (our)	PhysioData, Vision	2	**97.65±0.71**	**97.65±0.72**
Cimtay et al. [54]	EEG, Vision	4	53.80	–
Siddharth et al. et al. [52]	EEG, Vision	4	54.22	31.00
Lee et al. [55]	EEG, Vision	4	83.20	84.10
CLMER (our)	PhysioData, Vision	4	**96.47±1.29**	**96.47±1.28**
Y. Wu et al. [56]	EEG, Vision	9	95.12	94.22
CLMER (our)	PhysioData, Vision	9	**96.09±0.55**	**96.09±0.55**

TABLE II: Performance of different models on AMIGOS short dataset in terms of multi-class average accuracy (*Acc%*) and multi-class average F1 score (*F1%*) with stand deviations. V stands for Valence. A stands for Arousal. VA represents Valence-Arousal. The best results are highlighted in **bold**.

Method	Modality	Class	Result	
			Acc	*F1*
Strizhkova et al. [58]	PhysioData, Vision	V(2)	–	66.00±4
		A(2)	–	58.00±6
Kamran et al. [59]	PhysioData, Vision	V(2)	71.48±4.97	–
		A(2)	73.02±3.98	–
Siddharth et al. [52]	PhysioData, Vision	V(2)	78.23	74.00
		A(2)	81.47	72.00
Siddharth et al. [52]	Biosensing	V(2)	83.94	82.00
		A(2)	82.76	76.00
Menon et al. [60]	EEG, ECG, EDA	V(2)	87.10	–
		A(2)	80.50	–
CLMER(our)	PhysioData, Vision	V(2)	**88.37±0.61**	**88.37±0.58**
		A(2)	**83.56±0.97**	**83.53±0.96**
CLMER(our)	PhysioData, Vision	VA(4)	**81.42±1.19**	**81.41±1.16**

TABLE III: Performance of different models on MAN-II in terms of multi-class average accuracy (*Acc%*) and multi-class average F1 score (*F1%*) with stand deviations. The best result is highlighted in **bold**.

Method	Modality	Class	Result	
			Acc	*F1*
Conformer [26]	EEG	4	93.26	93.00
FetchEEG [26]	EEG	4	96.51	96.00
MobileVIT_s Pretrain [61]	EEG+Vision	4	95.27±1.14	95.28±1.13
MulT [29]	EEG+Vision	4	96.82	96.82
Husformer [30]	EEG+Vision	4	97.05	97.06
CLMER (our)	EEG+Vision	4	**97.75±1.63**	**97.75±1.62**

Experimental results demonstrate that the feature extraction module in CLMER plays a crucial role, leading to a significant improvement over existing methods. This further validates that CMLER is likely to exhibit greater stability and superior performance when applied to more complex real-world emotion recognition scenarios.

E. Ablation Study for Modality and CL Module

We evaluate the contributions of each component in our proposed Contrastive Learning based Multi-modal Emotion Recognition (CLMER) model, focusing on both modality utilization and the effectiveness of the contrastive module. Ablation experiments were conducted using the publicly available DEAP dataset [41], and AMIGOS [44], which are widely recognized for their comprehensive emotional data. We also use our private dataset MAN-II for ablation. The effectiveness of the ablation experiments was evaluated based on the accuracy and F1 score of the aforementioned nine-class and four-class classification tasks (see Section IV-A). In our initial exploration, we trained a 10-layer transformer network independently on three distinct modalities: preprocessed EEG signals, 3D landmark data, and AUs extracted from video recordings. Subsequently, we proceeded to combine the data from the three modalities pairwise to train the cross-modal attention based model employed in our network architecture.

We investigated three combinations: the pairing of EEG data with 3D Landmark information, the integration of EEG signals with AUs, and the combination of 3D Landmark data with AUs. By assessing each of above scenarios, we aimed to gain a deeper understanding of how each modality contributes to the collective performance of the model. In our proposed CLMER model, we employ the self-supervised contrastive learning techniques, to extract features for the classification task. Following this feature extraction phase, the resultant representations were subsequently fed into the fusion module to perform the task. This two-step approach not only enhances the ability of the model to capture intricate correlations between modalities but also improves the overall accuracy by providing comprehensive information for emotion recognition. By systematically integrating contrastive learning with a pairwise fusion module, our methodology seeks to optimize the effectiveness of multi-modal data utilization.

The experimental results of DEAP revealed that relying on a single modality provides an overly one-sided representation of human emotional states, rendering it insufficient for handling potentially complex emotion classification tasks, as evidenced by relatively low accuracy and F1 scores. When combining modality data in pairs, a significant improvement in training performance was observed (see Table IV). Notably, the fusion of the physiological signal modality with any visual information modality demonstrated remarkable enhancements, achieving an overall performance increase of nearly 50% compared to single-modality approaches. After integrating data from three modalities, both accuracy and F1 scores reached 94.91%, indicating that the multiple modalities enables the model to more comprehensively capture emotional states. Compared to single-modality approaches, multi-modal fusion clearly offers distinct advantages, achieving strong performance in the more complex nine-class classification task. With the assistance of contrastive learning (CL), the accuracy and F1 score further increased to 96.09%, demonstrating exceptional performance. Additionally, it is evident that the proposed model exhibits robust stability when addressing the samples' imbalance among class. The confusion matrix for experiments on the DEAP dataset (see Figure 4) exhibits a dark main diagonal with very light colors in other positions, indicating strong classification performance, minimal misclassification, and robust recognition capability of the model. The AMIGOS

TABLE IV: Performance of different mode on DEAP dataset in terms of nine-class average accuracy (*Acc%*) and nine-class average F1 score (*F1%*) with stand deviations. w/o: without, w/: with. The best result is highlighted in **bold**.

Mode	Modality			Result	
	EEG	*3D Landmark*	*AU*	*Acc*	*F1*
Single	✓	×	×	43.01±3.89	42.5±4.30
Single	×	✓	×	47.66±1.96	46.30±1.91
Single	×	×	✓	54.77±0.73	54.32±0.73
Dual	✓	✓	×	81.87±1.26	81.92±1.25
Dual	✓	×	✓	80.34±2.59	80.38±2.61
Dual	×	✓	✓	61.64±1.62	61.41±1.65
Fusion w/o CL	✓	✓	✓	94.91±0.89	94.91±0.89
Fusion w/ CL(our)	✓	✓	✓	**96.09±0.55**	**96.09±0.55**

dataset for ablation study is divided into two experimental settings: long-video stimuli and short-video stimuli, serving as complementary public datasets. We conducted comparative experiments across single-modality, dual-modality, and multi-modality approaches, as well as multi-modality with enhanced feature extraction. Overall, the results for long videos are better than those for short videos. This is in part because the long-video dataset contains a larger quantity of data with the same label, providing richer information. In addition, long videos create a more immersive emotional state for subjects, leading to improved data quality. The results (see Table V & Table VI) reveal that as the number of introduced modalities increases, both accuracy and F1 score show a gradual improvement. This trend aligns with the results observed on the DEAP dataset, further corroborating the importance of multi-modal fusion in emotion recognition. The confusion matrix for experiments on the AMIGOS dataset also show great performance with dark diagonal line (see Figure 5).

After incorporating contrastive learning for feature extraction, there was a noticeable improvement in short video

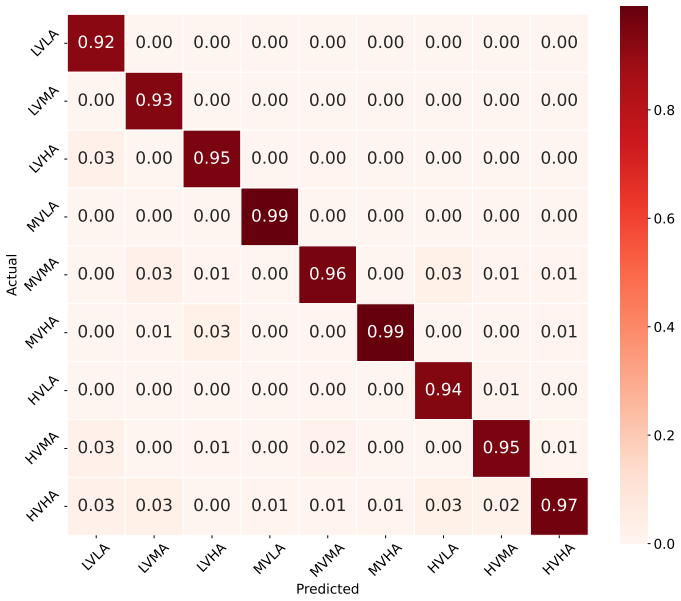


Fig. 4: Confusion matrix of CLMER on DEAP dataset

data, which had a lower accuracy and F1 scores before. The accuracy increased from 68.09% and F1 score of 68.07% to 81.96% accuracy and 81.94% F1 score with self-supervised feature extraction. The improvement of contrastive learning in long-video data, which already showed good fusion results, is less pronounced. However, it can be observed that the standard deviation of accuracy is reduced by approximately 36% with CL from 1.60 to 1.01. This indicates that after feature extraction using CL, the overall stability of models is enhanced. The similar situation can also be observed in previous experiments based on DEAP. For our privately orga-

TABLE V: Performance of different models on short-video AMIGOS in terms of four-class average accuracy (*Acc%*) and four-class average F1 score (*F1%*) with stand deviations. w/o: without, w/: with. The best result is highlighted in **bold**.

Mode	Modality			Result	
	EEG	*3D Landmark*	*AU*	*Acc*	*F1*
Single	✓	×	×	40.69±0.81	39.82±0.93
Single	×	✓	×	54.44±1.35	54.17±1.44
Single	×	×	✓	55.78±1.42	55.64±1.43
Dual	✓	✓	×	54.53±1.83	54.45±1.84
Dual	✓	×	✓	50.60±0.99	50.32±1.02
Dual	×	✓	✓	65.57±0.82	65.49±0.83
Fusion w/o CL	✓	✓	✓	68.09±0.40	68.07±0.39
Fusion w/ CL(our)	✓	✓	✓	**81.96±1.57**	**81.94±1.55**

nized dataset, MAN-II, the experimental results clearly show that the AU data in the serialized visual modality performs well in the single-modality setting and plays an important role in subsequent dual-modality and fusion processes. After the fusion of EEG and 3D Landmark with AU, the accuracy improved significantly from 41.72% and 50.11% to 84.79% and 89.39%, bringing nearly double the improvements, respectively. The result of 96.43% in the fusion mode demonstrates the effectiveness of fusing three modalities, while the use of CL for feature extraction further increased the accuracy

TABLE VI: Performance of different models on long-video AMIGOS in terms of four-class average accuracy (*Acc%*) and four-class average F1 score (*F1%*) with stand deviations. w/o: without, w/: with. The best result is highlighted in **bold**.

Mode	Modality			Result	
	EEG	*3D Landmark*	*AU*	*Acc*	*F1*
Single	✓	×	×	64.78±1.32	64.93±1.30
Single	×	✓	×	60.57±2.41	58.08±3.19
Single	×	×	✓	75.21±2.14	75.39±2.15
Dual	✓	✓	×	86.57±3.07	86.59±3.14
Dual	✓	×	✓	82.26±1.67	82.35±1.62
Dual	×	✓	✓	84.07±1.04	84.17±1.01
Fusion w/o CL	✓	✓	✓	90.07±1.60	90.09±1.58
Fusion w/ CL(our)	✓	✓	✓	**90.83±1.01**	**90.82±1.03**

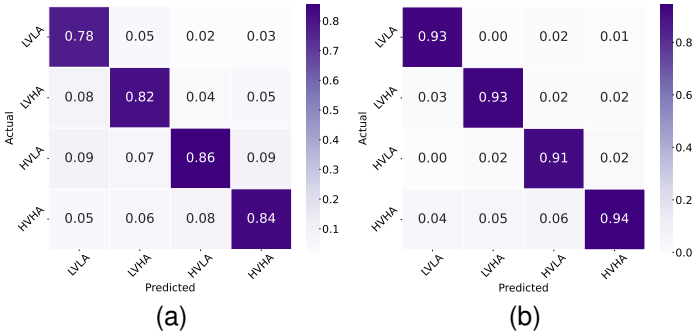


Fig. 5: Confusion matrix of CLMER on (a) AMIGOS Short and (b) AMIGOS Long datasets.

to 97.75% (see Table VII). The confusion matrix of MAN-II from the experimental results further supports the aforementioned conclusions, demonstrating excellent classification performance, with colors predominantly concentrated along the main diagonal (see Figure 6).

TABLE VII: Performance of different models on MAN-II in terms of four-class average accuracy (*Acc%*) and four-class average F1 score (*F1%*) with stand deviations. w/o: without, w/: with. The best result is highlighted in **bold**.

Mode	Modality			Result	
	EEG	*3D Landmark*	*AU*	*Acc*	*F1*
Single	✓	×	×	41.72±0.84	41.17±0.97
Single	×	✓	×	50.11±0.87	48.24±1.54
Single	×	×	✓	74.96±1.05	75.03±0.77
Dual	✓	✓	×	42.76±2.81	42.79±2.73
Dual	✓	×	✓	84.79±0.35	84.80±0.35
Dual	×	✓	✓	89.39±0.42	89.38±0.42
Fusion w/o CL	✓	✓	✓	96.43±1.91	96.44±1.94
Fusion w/ CL(our)	✓	✓	✓	**97.75±1.63**	**97.75±1.62**

This ablation study provides a detailed view of how different modalities and contrastive learning contribute to the CLMER framework. When using single modalities, the performance remains relatively limited: EEG alone captures intrinsic neural responses linked to emotions, but its discriminative power is reduced by noise and variability across subjects; 3D landmarks describe geometric facial movements, and AUs reflect muscle activations, yet both are subject to intentional masking or

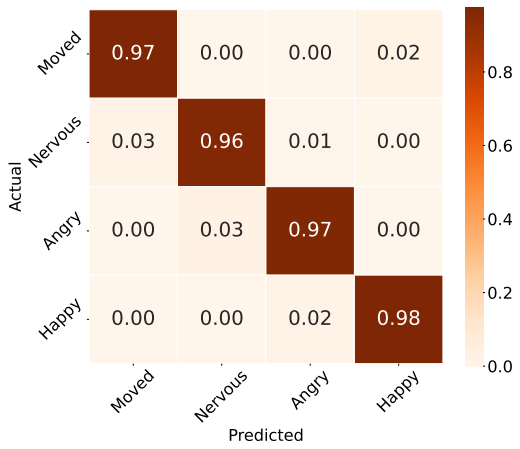


Fig. 6: Confusion matrix of CLMER on MAN-II dataset.

subtle individual differences. This explains why unimodal models in our experiments achieved notably lower accuracy and F1 scores compared to multimodal settings. When modalities are fused, the benefits become evident. Pairwise integration, such as EEG with 3D landmarks or EEG with AUs, already yields significant improvements over unimodal baselines. This demonstrates the complementary nature of internal physiological signals and external visual cues. EEG provides reliable and involuntary information about emotional states, while the serialized visual modality contributes expressive and interpretable external manifestations. The three-modality configuration (EEG + 3D landmarks + AUs) further enhances performance, achieving the best results across tasks and datasets. These outcomes suggest that each modality contributes non-redundant information and that their integration enables the framework to capture emotions from multiple perspectives, thereby improving robustness and generalization. The role of contrastive learning is equally critical. Adding the contrastive module consistently improved classification performance, increasing both accuracy and F1 scores beyond multimodal fusion alone. More importantly, it reduced performance variance across repeated runs, indicating greater stability. By maximizing agreement between augmented views of the same sample while separating different samples, contrastive learning extracts modality-specific but semantically aligned temporal representations. This not only bridges the structural gap between EEG and serialized visual data but also enhances discriminability in the fused space.

F. Ablation Study for Serialization

The efficiency of the CLMER framework is attributable to several design choices, one of which is the use of a serialization module that transforms facial images into sequential representations, replacing raw pixel-level inputs in the multimodal fusion process. To evaluate the effectiveness of the serialization strategy within CLMER, we conducted ablation studies by comparing it with two alternative methods. The first baseline employs a downsampling approach to resize captured facial images to 32×32×3, serving as the visual input modality. The second utilizes a pre-trained MobileViT [61] model to extract feature representations from the original 112×112×3 facial images, resulting in a 16-patch × 640-dimensional sequence. In both cases, temporal alignment with physiological signals is maintained at a one-second resolution. Experimental configurations are kept consistent with those in the main comparative experiments. To comprehensively evaluate the proposed method, we conduct experiments on two public datasets: DEAP and AMIGOS (short videos). Specifically, the DEAP dataset is used for binary, four-class, and nine-class emotion classification tasks, while the AMIGOS dataset focuses on binary and four-class tasks. To ensure a fair and informative comparison, model performance is assessed from three perspectives: memory consumption (i.e., GPU usage), training time, and classification effectiveness on the test set, measured by accuracy and F1-score (see Table VIII, Table IX and Table X).

TABLE VIII: Performance of different methods on DEAP dataset in terms of mulit-class accuracy (Acc%) and F1 score (F1%). The best result is highlighted in bold.

	Downsampling		Mobilevit_s		Serialization(3)	
	Acc	*F1*	*Acc*	*F1*	*Acc*	*F1*
VA (9)	81.21±2.77	81.19±2.75	87.58±2.98	87.58±2.99	**94.91±0.89**	**94.91±0.89**
VA (4)	76.43±4.16	76.28±4.22	88.90±2.11	87.76±2.28	**96.18±0.65**	**96.18±0.65**
Valence (2)	79.91±2.74	79.95±2.73	89.54±2.02	89.55±2.01	**95.49±0.48**	**95.49±0.48**
Arousal (2)	82.22±0.55	82.16±0.56	87.79±1.75	87.81±1.77	**95.05±1.28**	**95.03±1.27**

TABLE IX: Performance of different methods on AMIOGS Short dataset in terms of mulit-class accuracy (Acc%) and F1 score (F1%). The best result is highlighted in bold.

	Downsampling		Mobilevit_s		Serialization(3)	
	Acc	*F1*	*Acc*	*F1*	*Acc*	*F1*
VA (4)	31.06±1.62	21.44±3.69	28.34±1.10	24.43±1.22	**68.09±0.40**	**68.07±0.39**
Valence (2)	62.43±0.50	48.05±0.70	60.48±2.39	50.12±3.24	**78.29±1.99**	**77.98±2.01**
Arousal (2)	55.90±1.61	41.36±1.17	55.58±1.57	45.18±4.98	**69.61±1.39**	**69.48±1.52**

To facilitate comparison, we used the mean value of each metric as the reference unit (set to 1) and computed the ratio of each method to this mean. The reciprocals of these ratios were then taken to represent time efficiency and space efficiency, so that larger values indicate better performance. This inversion ensures that shorter training times and lower GPU memory consumption correspond to higher efficiency scores, which can be interpreted as capability values for comparison. For accuracy, to better evaluate the performance of each method, we first computed the ratio relative to the mean, treating the mean as the unit value of 1. We then calculated the difference between each method and the mean, and applied z-score normalization to standardize these differences to have a mean of zero. Since the resulting standardized values, following a normal distribution $N(0, 1)$, may include negative numbers, we shifted the axis by two units to the right to represent the capability of each method in terms of training accuracy. The performance of three methods based on the above procedure is shown in the Figure 7 and 8.

From the radar charts, it can be observed that the Downsampling method consistently performs the worst across all three metrics, remaining below the average level of the three approaches. The MobileViT_s-based multimodal fusion method

TABLE X: The average time consumption, space consumption, and accuracy across the three methods.

	DEAP			AMIGOS		
	Downsampling	Mobilevit_s	Serialization	Downsampling	Mobilevit_s	Serialization
GPU Memory(MiB)	21728	2534	2844	21730	2148	2856
Time(s)	19084	6820	15226	20026	7097	16496
Accuracy(%)	79.94	88.45	95.41	49.80	48.13	72.00

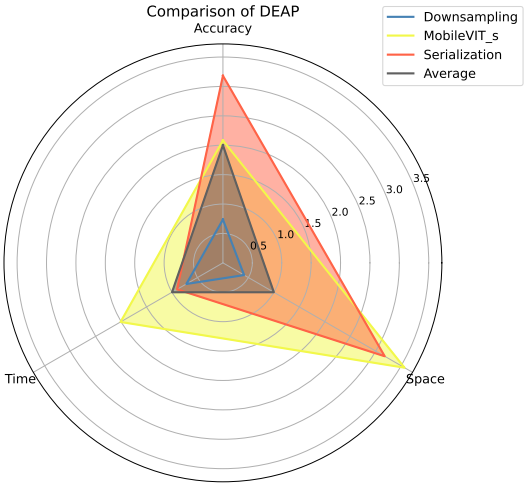


Fig. 7

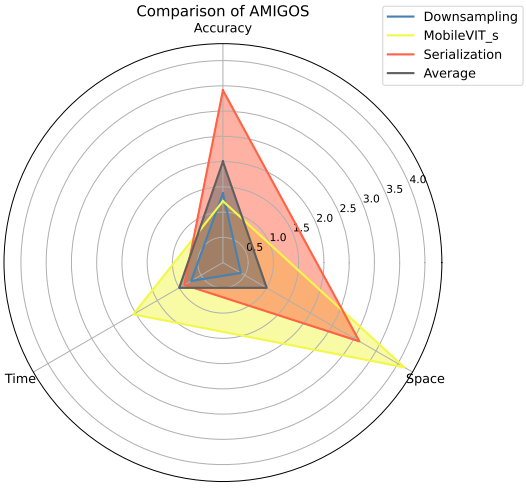


Fig. 8

preprocesses the image modality using a pre-trained model and incorporates it with physiological signals during training, resulting in only two modalities and thus two times cross-modal attention operations. This design yields clear advantages in both training time and memory consumption compared to the other methods, although its overall performance exceeds that of Downsampling, it's accuracy remains only around the average on the DEAP dataset and even falls below Downsampling on the AMIGOS dataset.

In contrast, the proposed serialization method transforms the original image modality into two distinct modalities during computation. Combined with the physiological signal modality, this forms a three-modality framework and entails six times cross-modal attention operations in total. While this increases the number of modalities and extends training time compared to the MobileViT_s-based method, it still achieves faster training than the Downsampling approach and maintains comparable memory consumption to MobileViT_s. More importantly, serialization delivers a substantial accuracy gain, as it exceeds the average accuracy on the DEAP dataset by 7.48 percentage points and surpassing MobileViT_s by 6.96 percentage points. On the AMIGOS dataset, it outperforms the second-best method (Downsampling) by 22.2 percentage points and 23.87 comparing to MobileVit_s, demonstrating a significant and consistent advantage in classification performance. The longer training time of serialization is partly attributable to computations being performed on an NVIDIA RTX 3090, which lags behind state-of-the-art hardware; on more advanced devices, this time gap would be considerably reduced, making the significant accuracy advantage of serialization even more compelling.

G. Parameter Analysis

In the multi-modal fusion module, a cross-modal attention mechanism is used to perform computation between data between every two modalities. Before performing this operation, there is a procedure unifying the tokens length to simplify the computation process. The token length significantly influences the performance of fusion module. We conducted an experiment to determine the optimal token length on two-class and four-class tasks on both DEAP and short videos' AMIGOS dataset and additional nine-class task on DEAP. In the experiment, we keep other experimental settings such as equipment, seeds, and dataset structure consistent. Changing the length of the token multiple times, we obtained the experimental results shown in Table VIII and Table IX.

For DEAP, it can be observed that when token length is around 30 to 40, training performance is generally suboptimal. However, as the length exceeds the maximum length 40 initially provided by three modalities, the performance gradually improves, reaching its optimal level around 60 for two-class tasks, approximately 70 for four-class task and 50 for nine-class task (see Table XI and Figure 7). Optimal token length slightly longer than the 40-length channels data provided by the physiological data, and much longer than the 3-channel and 2-channel data provided by the other two modalities. We concluded that a token length that is too short may limit the performance of the long-token modality, while an excessively long token may overly dilute the modality data density. Therefore, when determining the length of token for fusion using cross-modal attention, it is essential to avoid the excessive compression of long-token modalities, and it is equally important to prevent the excessive dilution of short-token modalities. The experiments conducted on the AMIGOS Short dataset, as illustrated in Table XII and Figure 8. The number of channels provided by the physiological data decreased to 17, while the other two modalities remained unchanged. It is evident that the optimal fusion token length is

TABLE XI: Performance of different token length on DEAP dataset in terms of mulit-class accuracy (*Acc%*) and F1 score (*F1%*). The best result is highlighted in **bold**.

Token length	Valence (2)		Arousal (2)		VA (4)		VA (9)	
	Acc	*F1*	*Acc*	*F1*	*Acc*	*F1*	*Acc*	*F1*
30	94.61	94.62	94.22	94.24	92.72	92.72	94.44	94.45
40	92.22	92.23	92.89	92.89	94.50	94.50	96.50	96.50
50	96.56	96.56	95.28	95.28	95.61	95.62	**97.01**	**97.02**
60	**98.11**	**98.11**	**96.83**	**96.84**	97.22	97.22	96.22	96.22
70	96.22	96.23	95.56	95.56	**97.44**	**97.44**	96.89	96.88

TABLE XII: Performance of different token length on AMIGOS Short dataset in terms of mulit-class accuracy (*Acc%*) and F1 score (*F1%*). The best result is highlighted in **bold**.

Token length	Valence (2)		Arousal (2)		VA (4)	
	Acc	*F1*	*Acc*	*F1*	*Acc*	*F1*
5	84.69	84.67	78.57	78.52	68.82	68.77
10	87.61	87.62	**84.45**	**84.42**	82.79	82.77
15	**89.26**	**89.26**	82.46	82.43	**83.82**	**83.80**
20	87.54	87.52	81.36	81.30	82.13	82.12
25	87.26	87.24	80.55	80.46	80.48	80.45

located from 10 to 20, a range that correspond to the maximum token length provided, while ideally remaining as close as possible to the two smaller modalities. It is important to avoid using excessively short token lengths. As shown in the figure, when a token length of 5 is used, the training convergence speed significantly decreases, and the final performance is considerably lower than that of other token lengths. This further emphasizes the necessity of prioritizing the longest token length provided by all modalities when performing fusion using cross-modal attention.

V. Conclusion

In this paper, we propose CLMER, a contrastive learning based multi-modal cross-attention framework designed to address potential real-world demands for complex emotion recognition. CLMER applies a serialization operation to visual image data, which reduces modality discrepancies, unifies modality structures, and simultaneously simplifies the overall architecture of the multi-modal fusion model training framework. Contrastive learning is introduced as a feature extraction method to further enhance the effectiveness of modality fusion. Serialization procedure not only eliminates the need for modality-specific feature extraction approaches but also provides solid alignment for the fusion process. Moreover, by compressing the scale of the visual data, it significantly reduces the resource requirements for training, including memory consumption, while substantially enhancing computational efficiency. The experimental results clearly demonstrate that multi-modal fusion outperforms single-modality approaches in tackling more complex emotion recognition tasks on the two public datasets, DEAP and AMIGOS, as well as the expanded private dataset, MAN-II. We attribute this improvement to the powerful modality fusion capabilities of the CLMER framework, as well as the comprehensive internal and external data support. Unlike traditional feature extraction methods,

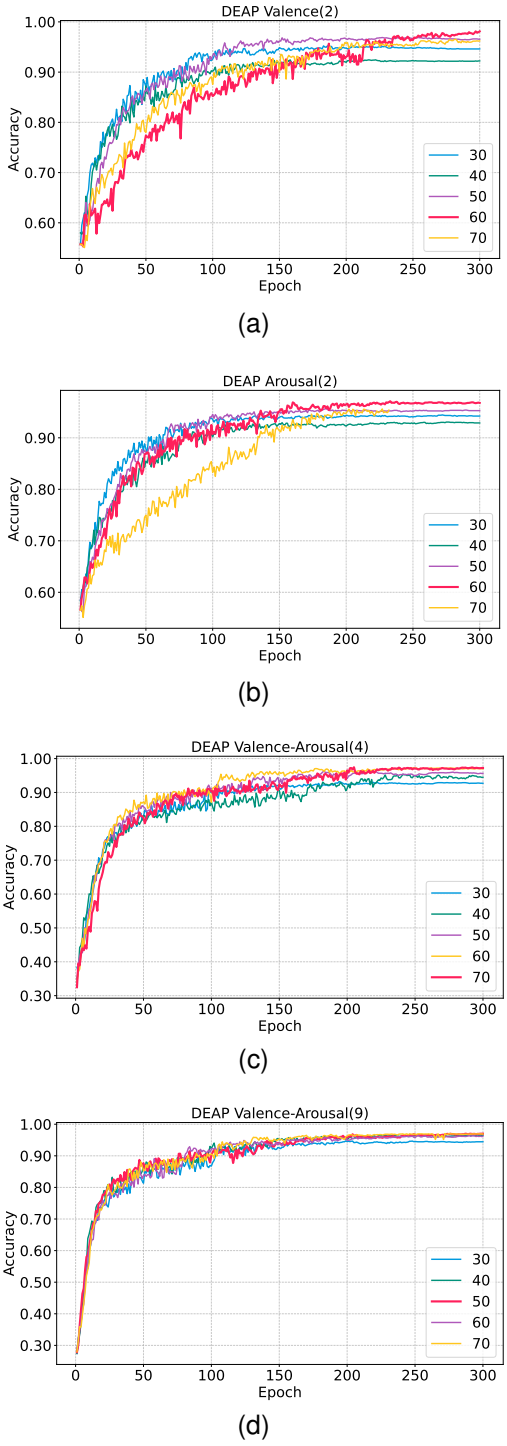
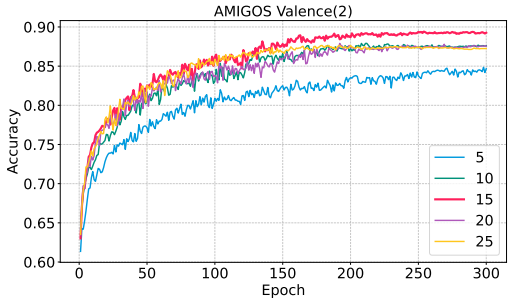
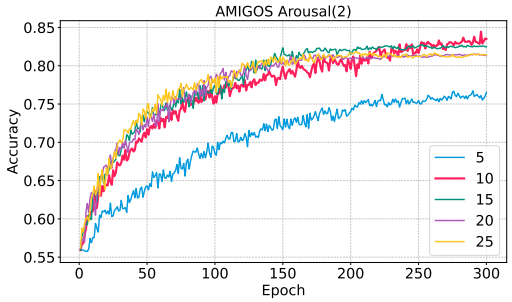


Fig. 9: Accuracy of CLMER with different token length on DEAP dataset. (a) illustrates the binary classification task on the valence dimension. (b) illustrates the binary classification task on the arousal. (c) illustrates the four-class classification task on the valence-arousal. (d) illustrates the nine-class classification task on the valence-arousal. The curves of best results are highlighted in red.

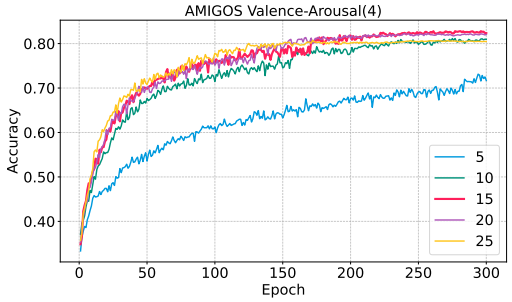
CLMER employs contrastive learning to derive representations from sequential data, facilitating subsequent multi-modal fusion. While achieving outstanding performance comparable



(a)



(b)



(c)

Fig. 10: Accuracy of CLMER with different token length on AMIGOS Short dataset. The best results are highlighted in **bold** and red. (a) illustrates the binary classification task on the valence dimension. (b) illustrates the binary classification task on the arousal dimension. (c) illustrates the four-class classification task on the valence-arousal dimensions.

to other high-performance models for emotion recognition, CLMER is also exhibiting robust generalization and stability in emotion recognition tasks.

REFERENCES

[1] S. K. Khare, V. Blanes-Vidal, E. S. Nadimi, and U. R. Acharya, "Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations," *Information fusion*, vol. 102, p. 102019, 2024.

[2] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0893608014002135

[3] K. Sarvakar, R. Senkamalavalli, S. Raghavendra, J. S. Kumar, R. Manjunath, and S. Jaiswal, "Facial emotion recognition using convolutional neural networks," *Materials Today: Proceedings*, vol. 80, pp. 3560–3564, 2023.

[4] J. Wei, G. Hu, X. Yang, A. T. Luu, and Y. Dong, "Learning facial expression and body gesture visual information for video emotion recognition," *Expert Systems with Applications*, vol. 237, p. 121419, 2024.

[5] M. J. Al-Dujaili and A. Ebrahimi-Moghadam, "Speech emotion recognition: a comprehensive survey," *Wireless Personal Communications*, vol. 129, no. 4, pp. 2525–2561, 2023.

[6] M. Egger, M. Ley, and S. Hanke, "Emotion recognition from physiological signal analysis: A review," *Electronic Notes in Theoretical Computer Science*, vol. 343, pp. 35–55, 2019.

[7] H. Yang, J. Han, and K. Min, "A multi-column cnn model for emotion recognition from eeg signals," *Sensors*, vol. 19, no. 21, p. 4736, 2019.

[8] T. Fan, S. Qiu, Z. Wang, H. Zhao, J. Jiang, Y. Wang, J. Xu, T. Sun, and N. Jiang, "A new deep convolutional neural network incorporating attentional mechanisms for ecg emotion recognition," *Computers in Biology and Medicine*, vol. 159, p. 106938, 2023.

[9] S. Dutta, B. K. Mishra, A. Mitra, and A. Chakraborty, "An analysis of emotion recognition based on gsr signal," *ECS Transactions*, vol. 107, no. 1, p. 12535, 2022.

[10] J. Zhang, Z. Yin, P. Chen, and S. Nichele, "Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review," *Information Fusion*, vol. 59, pp. 103–126, 2020.

[11] W. Liu, J.-L. Qiu, W.-L. Zheng, and B.-L. Lu, "Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 14, no. 2, pp. 715–729, 2021.

[12] Y. Zhang, X. Ding, K. Gong, Y. Ge, Y. Shan, and X. Yue, "Multimodal pathway: Improve transformers with irrelevant data from other modalities," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 6108–6117.

[13] P. P. Liang, A. Zadeh, and L.-P. Morency, "Foundations and trends in multimodal machine learning: Principles, challenges, and open questions," *arXiv preprint arXiv:2209.03430*, 2022.

[14] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[15] S. Siriwardhana, A. Reis, R. Weerasekera, and S. Nanayakkara, "Jointly fine-tuning" bert-like" self supervised models to improve multimodal speech emotion recognition," *arXiv preprint arXiv:2008.06682*, 2020.

[16] Y. Khaireddin and Z. Chen, "Facial emotion recognition: State of the art performance on fer2013," *arXiv preprint arXiv:2105.03588*, 2021.

[17] A. Zadeh, Y. Chong Lim, T. Baltrusaitis, and L.-P. Morency, "Convolutional experts constrained local model for 3d facial landmark detection," in *Proceedings of the IEEE international conference on computer vision workshops*, 2017, pp. 2519–2528.

[18] T. Baltrušaitis, M. Mahmoud, and P. Robinson, "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. IEEE, 2015, pp. 1–6.

[19] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[20] ——, "Facial expression and emotion." *American psychologist*, vol. 48, no. 4, p. 384, 1993.

[21] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[22] D. Issa, M. F. Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," *Biomedical Signal Processing and Control*, vol. 59, p. 101894, 2020.

[23] A. F. Adoma, N.-M. Henry, and W. Chen, "Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition," in *2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*. IEEE, 2020, pp. 117–121.

[24] A. Saxena, A. Khanna, and D. Gupta, "Emotion recognition and detection methods: A comprehensive survey," *Journal of Artificial Intelligence and Systems*, vol. 2, no. 1, pp. 53–79, 2020.

[25] S. M. Alarcao and M. J. Fonseca, "Emotions recognition using eeg signals: A survey," *IEEE transactions on affective computing*, vol. 10, no. 3, pp. 374–393, 2017.

[26] Y. Liang, C. Zhang, S. An, Z. Wang, K. Shi, T. Peng, Y. Ma, X. Xie, J. He, and K. Zheng, "Fetcheeg: a hybrid approach combining feature extraction and temporal-channel joint attention for eeg-based emotion classification," *Journal of Neural Engineering*, vol. 21, no. 3, p. 036011, 2024.

[27] N. Jain, S. Kumar, A. Kumar, P. Shamsolmoali, and M. Zareapoor, "Hybrid deep neural networks for face emotion recognition," *Pattern Recognition Letters*, vol. 115, pp. 101–106, 2018.

[28] S. M. S. A. Abdullah, S. Y. A. Ameen, M. A. Sadeeq, and S. Zeebaree, "Multimodal emotion recognition using deep learning," *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, pp. 73–79, 2021.

[29] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the conference. Association for computational linguistics. Meeting*, vol. 2019. NIH Public Access, 2019, p. 6558.

[30] R. Wang, W. Jo, D. Zhao, W. Wang, A. Gupte, B. Yang, G. Chen, and B.-C. Min, "Husformer: A multi-modal transformer for multi-modal human state recognition," *IEEE Transactions on Cognitive and Developmental Systems*, 2024.

[31] S. Koelstra and I. Patras, "Fusion of facial expressions and eeg for implicit affective tagging," *Image and Vision Computing*, vol. 31, no. 2, pp. 164–174, 2013.

[32] I. Hosseini, M. Z. Hossain, Y. Zhang, and S. Rahman, "Deep learning model for simultaneous recognition of quantitative and qualitative emotion using visual and bio-sensing data," *Computer Vision and Image Understanding*, vol. 248, p. 104121, 2024.

[33] A. Mehrabian, "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament," *Current Psychology*, vol. 14, pp. 261–292, 1996.

[34] P. Bachman, R. D. Hjelm, and W. Buchwalter, "Learning representations by maximizing mutual information across views," *Advances in neural information processing systems*, vol. 32, 2019.

[35] J. Liu and S. Chen, "Timesurl: Self-supervised contrastive learning for universal time series representation learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 12, 2024, pp. 13 918–13 926.

[36] E. Eldele, M. Ragab, Z. Chen, M. Wu, C.-K. Kwoh, X. Li, and C. Guan, "Self-supervised contrastive representation learning for semi-supervised time-series classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[37] B. Tadas, Z. Amir, L. Y. Chong, and L.-M. Philippe, "Openface 2.0: Facial behavior analysis toolkit," in *13th IEEE International Conference on Automatic Face & Gesture Recognition*, 2018.

[38] T. Chen, Y. Sun, Y. Shi, and L. Hong, "On sampling strategies for neural network-based collaborative filtering," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017, pp. 767–776.

[39] M. N. Mohsenvand, M. R. Izadi, and P. Maes, "Contrastive representation learning for electroencephalogram classification," in *Machine Learning for Health*. PMLR, 2020, pp. 238–253.

[40] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[41] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis; using physiological signals," *IEEE transactions on affective computing*, vol. 3, no. 1, pp. 18–31, 2011.

[42] T. Baltrusaitis, P. Robinson, and L.-P. Morency, "Constrained local neural fields for robust facial landmark detection in the wild," in *Proceedings of the IEEE international conference on computer vision workshops*, 2013, pp. 354–361.

[43] R. Plutchik, "The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice," *American scientist*, vol. 89, no. 4, pp. 344–350, 2001.

[44] J. A. Miranda-Correa, M. K. Abadi, N. Sebe, and I. Patras, "Amigos: A dataset for affect, personality and mood research on individuals and groups," *IEEE transactions on affective computing*, vol. 12, no. 2, pp. 479–493, 2018.

[45] R. R. McCrae and O. P. John, "An introduction to the five-factor model and its applications," *Journal of personality*, vol. 60, no. 2, pp. 175–215, 1992.

[46] D. Watson, L. A. Clark, and A. Tellegen, "Development and validation of brief measures of positive and negative affect: the panas scales." *Journal of personality and social psychology*, vol. 54, no. 6, p. 1063, 1988.

[47] E. Wood, T. Baltrusaitis, X. Zhang, Y. Sugano, P. Robinson, and A. Bulling, "Rendering of eyes for eye-shape registration and gaze estimation," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3756–3764.

[48] Y.-C. Wu, L.-W. Chiu, C.-C. Lai, B.-F. Wu, and S. S. Lin, "Recognizing, fast and slow: Complex emotion recognition with facial expression detection and remote physiological measurement," *IEEE Transactions on Affective Computing*, vol. 14, no. 4, pp. 3177–3190, 2023.

[49] Y. Huang, J. Yang, S. Liu, and J. Pan, "Combining facial expressions and electroencephalography to enhance emotion recognition," *Future Internet*, vol. 11, no. 5, p. 105, 2019.

[50] Y. Zhao and D. Chen, "Expression eeg multimodal emotion recognition method based on the bidirectional lstm and attention mechanism," *Computational and Mathematical Methods in Medicine*, vol. 2021, no. 1, p. 9967592, 2021.

[51] Y. Zhang, C. Cheng, and Y. Zhang, "Multimodal emotion recognition using a hierarchical fusion convolutional neural network," *IEEE access*, vol. 9, pp. 7943–7951, 2021.

[52] T.-P. Jung, T. J. Sejnowski *et al.*, "Utilizing deep learning towards multi-modal bio-sensing and vision-based affective computing," *IEEE Transactions on Affective Computing*, vol. 13, no. 1, pp. 96–107, 2019.

[53] S. Wang, J. Qu, Y. Zhang, and Y. Zhang, "Multimodal emotion recognition from eeg signals and facial expressions," *IEEE Access*, vol. 11, pp. 33 061–33 068, 2023.

[54] Y. Cimtay, E. Ekmekcioglu, and S. Caglar-Ozhan, "Cross-subject multimodal emotion recognition based on hybrid fusion," *IEEE Access*, vol. 8, pp. 168 865–168 878, 2020.

[55] J.-H. Lee, J.-Y. Kim, and H.-G. Kim, "Emotion recognition using eeg signals and audiovisual features with contrastive learning," *Bioengineering*, vol. 11, no. 10, p. 997, 2024.

[56] Y. Wu and J. Li, "Multi-modal emotion identification fusing facial expression and eeg," *Multimedia Tools and Applications*, vol. 82, no. 7, pp. 10 901–10 919, 2023.

[57] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[58] V. Strizhkova, H. Kachmar, H. Chaptoukaev, R. Kalandadze, N. Kukhilava, T. Tsmindashvili, N. Abo-Alzahab, M. A. Zuluaga, M. Balazia, A. Dantcheva *et al.*, "Mvp: Multimodal emotion recognition based on video and physiological signals," *arXiv preprint arXiv:2501.03103*, 2025.

[59] K. Ali and C. E. Hughes, "A unified biosensor–vision multi-modal transformer network for emotion recognition," *Biomedical Signal Processing and Control*, vol. 102, p. 107232, 2025.

[60] A. Menon, A. Natarajan, R. Agashe, D. Sun, M. Aristio, H. Liew, Y. S. Shao, and J. M. Rabaey, "Efficient emotion recognition using hyperdimensional computing with combinatorial channel encoding and cellular automata," *Brain informatics*, vol. 9, no. 1, p. 14, 2022.

[61] S. Mehta and M. Rastegari, "Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer," *arXiv preprint arXiv:2110.02178*, 2021.