bioengineering

MDPI

Article

Emotion Recognition Using EEG Signals and Audiovisual Features with Contrastive Learning

Ju-Hwan Lee [1], **Jin-Young Kim** [1] and **Hyoung-Gook Kim** [2,*]

[1] Department of Intelligent Electronics and Computer Engineering, Chonnam National University, 77 Yongbong-ro, Buk-gu, Gwangju 61186, Republic of Korea; juhwanlee@jnu.ac.kr (J.-H.L.); beyondi@jnu.ac.kr (J.-Y.K.)

[2] Department of Electronic Convergence Engineering, Kwangwoon University, 20 Gwangun-ro, Nowon-gu, Seoul 01897, Republic of Korea

* Correspondence: hkim@kw.ac.kr

Abstract: Multimodal emotion recognition has emerged as a promising approach to capture the complex nature of human emotions by integrating information from various sources such as physiological signals, visual behavioral cues, and audio-visual content. However, current methods often struggle with effectively processing redundant or conflicting information across modalities and may overlook implicit inter-modal correlations. To address these challenges, this paper presents a novel multimodal emotion recognition framework which integrates audio-visual features with viewers' EEG data to enhance emotion classification accuracy. The proposed approach employs modality-specific encoders to extract spatiotemporal features, which are then aligned through contrastive learning to capture inter-modal relationships. Additionally, cross-modal attention mechanisms are incorporated for effective feature fusion across modalities. The framework, comprising pre-training, fine-tuning, and testing phases, is evaluated on multiple datasets of emotional responses. The experimental results demonstrate that the proposed multimodal approach, which combines audio-visual features with EEG data, is highly effective in recognizing emotions, highlighting its potential for advancing emotion recognition systems.

Keywords: emotion recognition; multimodal learning; contrastive learning; cross-attention mechanism

check for updates

Citation: Lee, J.-H.; Kim, J.-Y.; Kim, H.-G. Emotion Recognition Using EEG Signals and Audiovisual Features with Contrastive Learning. *Bioengineering* **2024**, *11*, 997. https://doi.org/10.3390/bioengineering11100997

Academic Editor: Larbi Boubchir

Received: 10 September 2024
Revised: 28 September 2024
Accepted: 1 October 2024
Published: 3 October 2024

1. Introduction

Emotions are multifaceted psychological phenomena which result from the interaction of both internal cognitive states and external environmental stimuli. They encompass a wide range of physiological, behavioral, and subjective experiences which reflect an individual's response to a stimulus [1]. These stimuli can include sensory inputs from multimedia content, such as videos and images, which play a crucial role in evoking emotional responses. Platforms like YouTube, Netflix, and TikTok provide users with dynamic audiovisual content which not only conveys information but also serves as a significant medium for emotional engagement. These interactions highlight the complex nature of emotions, which are influenced by both personal dispositions and external influences. As a result, emotion recognition technology has gained importance in various applications, such as personalized content recommendation [2], therapeutic interventions in medical systems [3], and emotion-driven marketing strategies [4], helping to better understand and respond to user emotions in digital environments.

Conventional emotion recognition typically analyzes users' responses to stimuli, primarily utilizing physiological signal responses (e.g., EEG or ECG) [5,6] and visual behavioral responses (e.g., facial expressions and voice) [7,8]. Physiological signals offer the advantage of objectively capturing unconscious emotional reactions, while visual behavioral responses enable noninvasive and real-time analysis based on users' external reactions. An alternative approach to emotion recognition involves analyzing the stimulus itself and recognizing emotions by examining the characteristics of multimedia content [9,10]. The audio-visual signals in multimedia content provide powerful emotional stimuli. For instance, dark lighting and tense background music in horror movies induce fear and anxiety, while bright color schemes and upbeat music in comedy programs evoke joy and laughter. These intrinsic emotions can typically be inferred from visual information (color, brightness, and movement) and audio information (voice energy, frequency patterns, and volume), which is a widely recognized direct emotion recognition method in video emotion recognition [11–13].

Recently, multimodal emotion recognition [14,15], which allows for richer information utilization than stimulus- and response-based single-modality emotion recognition, has gained attention. This approach enables a more sophisticated understanding of emotional states by simultaneously analyzing multiple signals such as physiological responses, visual behavioral responses, and audio-visual signals. It has also been discovered that a more robust emotion recognition model can be acquired through the collaboration of different modalities [16,17]. However, multimodal approaches primarily focus on fusion at the feature level [18–23] and decision level [24–27], which can present challenges in processing redundant and conflicting information between modalities [28]. Additionally, concerns have been raised about potentially overlooking implicit correlations between modalities during the formation of high-dimensional feature vectors [29].

To address these issues, we propose an emotion recognition method which leverages contrastive learning [30] and cross-modal attention. Contrastive learning is a technique which clearly learns the relationships between high-dimensional data by placing similar data closer together and dissimilar data farther apart. Cross-modal attention enhances inter-modality interactions by selectively focusing on important information from different modalities. We propose a multimodal emotion recognition method which utilizes both the emotions inherent in video and audio stimuli and the physiological signals directly experienced by humans in response to these stimuli. For this, we employ audio-visual signals and EEG as physiological signals. EEG has proven to be a powerful tool for capturing changes in emotional states, recently demonstrating significant improvements in emotion recognition performance when combined with deep learning [31]. The choice of physiological signals over nonverbal cues is based on emotion theories [32,33]. Physiological signals, being unconscious bodily changes controlled by the autonomic nervous system, can potentially represent emotions more reliably than voluntary or involuntary facial behaviors.

The contributions of this paper can be summarized as follows:

1. In this paper, a multimodal emotion recognition framework is proposed based on audio-visual signals and EEG signals to consider both response and stimulus signals.
2. We integrate modality-specific networks and temporal convolutional networks (TCNs) into modal encoders to extract spatiotemporal representations of multimodal data while employing contrastive learning to capture intra-modal, inter-modal, and inter-class relationships in a shared embedding space.
3. We utilize cross-modal attention mechanisms to enhance the interactions between the extracted representations and to focus on the most salient information from each modality.
4. We demonstrate the superior performance of our proposed method in emotion recognition through benchmark datasets and our own collected dataset.

The remainder of this paper is organized as follows. Section 2 provides a review of the related work, the proposed method is explained in Section 3, experimental results validating the effectiveness of our emotion recognition approach are presented in Section 4, Section 5 discusses the limitations of this work, and finally, Section 6 concludes this paper and suggests directions for future research.

2. Related Work

In this section, we present a concise review of the relevant literature, focusing on recent advancements in the field of emotion recognition. Our discussion is structured around three key areas which form the foundation of our proposed framework: multimodal emotion recognition, contrastive learning, and cross-modal attention.

2.1. Multimodal Emotion Recognition

Emotion recognition has evolved significantly from its early focus on unimodal approaches. Initially, researchers explored individual modalities such as facial expression analysis [34], speech signal processing [35], and physiological signal analysis [36]. These unimodal methods provided valuable insights into specific aspects of emotion expression. For instance, EEG-based emotion recognition using hybrid CNN and LSTM models demonstrated promising results in capturing brain activity patterns associated with emotions [37].

However, emotion recognition systems that rely on a single modality often face challenges in real-world environments due to factors such as noise interference and signal degradation. To overcome these limitations, researchers have introduced multimodal emotion recognition techniques which integrate two or more modalities [38–40].

Among various multimodal approaches, the fusion of visual and audio data is particularly prevalent. This combination benefits from relatively straightforward data collection and provides complementary information. However, it is important to note that within the visual modality, dynamic stimuli (such as videos) often provide richer emotional information compared with static stimuli (like images). Dynamic visual content captures the temporal evolution of emotions, allowing for a more comprehensive representation of emotional states [41].

Furthermore, the impact of audio, video, and combined audio-video stimuli on emotional responses varies significantly. Audio stimuli can evoke emotions through tone, rhythm, pitch, and acoustic features, while video stimuli provide visual cues such as facial expressions, body language, color schemes, lighting, and movement patterns. When combined, these audio-visual signals in multimedia content serve as powerful emotional stimuli, creating a more immersive experience capable of inducing a wide range of stronger and more complex emotional responses in viewers than either modality alone.

Despite these advantages, multimodal approaches using only visual and auditory signals predominantly capture external emotional expressions, potentially overlooking essential aspects of a person’s internal emotional state. To address this limitation, some researchers have developed frameworks which merge visual or audio data with physiological signals [42,43]. This approach offers the capability to detect subtle or concealed emotions by leveraging both external and internal cues simultaneously.

Further expanding on this concept, some researchers have explored tri-modal systems which integrate visual, audio, and physiological signals simultaneously [44,45]. These approaches assess a more holistic range of emotional indicators, potentially enhancing the sensitivity and robustness of emotion recognition.

Understanding the distinctions between different modalities and the strengths of various modal combinations is crucial for developing more nuanced and effective emotion recognition systems which can adapt to different types of input and scenarios.

2.2. Contrastive Learning

Contrastive learning is a self-supervised learning paradigm where a model is trained to learn meaningful representations of input data by contrasting similar and dissimilar examples [46]. While initially developed for unimodal data, particularly in computer vision, its usage has recently expanded to multimodal learning. Contrastive learning has gained prominence in multimodal contexts for several key reasons. First, it is effective in aligning cross-modal data. It enables the mapping of data from different modalities into a shared embedding space. Second, it is highly data-efficient as it can derive valuable representations from unlabeled data, reducing the need for large labeled multimodal datasets and lowering associated costs and time investments. Third, it addresses challenges such as noise and domain bias [47].

Contrastive learning has also been explored in the context of emotion recognition within multimodal frameworks, though this remains an emerging area of study. Dissanayake et al. [48] introduced a method using wearable sensor data to learn representations for individual vital sign modalities via a self-supervised learning mechanism, which are subsequently fused for emotion recognition tasks. Jiang et al. [49] proposed a contrastive learning framework for emotion recognition based on EEG and eye movement data. Similarly, Tang et al. [50] employed a multimodal approach centered on physiological signals, designing emotion encoders which work across domains and modalities, and they introduced a hierarchical network to integrate these modalities in a structured manner. This multimodal contrastive learning approach leverages information from diverse modalities, such as visual, auditory, and EEG data, to enable more precise and robust classification of emotional states. By utilizing the unique emotional features inherent in each modality, it captures richer emotional expressions while compensating for the shortcomings of any single modality, making it more suitable for real-world applications.

However, to the best of our knowledge, no previous studies have been carried out using contrastive learning for emotion classification which combine audiovisual data (especially video) with EEG. We suppose this gap in the research is primarily due to the complexity of the stimuli, particularly when incorporating video content. This complexity presents significant challenges in applying contrastive learning techniques to such diverse and rich modalities simultaneously.

2.3. Cross-Modal Attention

Cross-modal attention is a crucial technique in multimodal learning that enables the effective modeling of relationships between different modalities. This mechanism learns how the features from one modality influence those in another, thereby enhancing the interaction and synergy across modalities. For instance, the authors of [51] employed a cross-modal transformer module to capture long-range dependencies between emotional cues and other modal elements in conversational contexts, emphasizing the interaction between text and audio modalities to adaptively promote convergence between them. Similarly, the authors of [52] applied a cross-modal attention mechanism within an audio-visual fusion model for emotion recognition, aiming to learn the correlations between fused feature representations and the representations of individual modalities.

In another approach, the authors of [53] introduced a multimodal method for music emotion recognition, utilizing a cross-modal attention mechanism to integrate information from various modalities into a hierarchical structure. This method efficiently combines the strengths of each modality. Furthermore, the authors of [40] proposed an end-to-end multimodal emotion recognition framework which leverages cross-modal attention to fuse audio and visual data. This approach capitalizes on the complementary properties of different modalities while maintaining modality-specific characteristics, resulting in more discriminative embeddings, more compact within-class representations, and greater separation between classes. Additionally, the authors of [54] adopted a cross-modal attention mechanism for multidimensional emotion recognition, effectively modeling the relationships between audio, visual, and textual modalities. By simultaneously capturing both intra-modality and inter-modality interactions, this approach produces a more refined and sophisticated feature representation.

3. Proposed Method

In this paper, we present a framework to enhance emotion recognition by leveraging multimodal input data, including visual, audio, and EEG signals. Our approach combines multimodal contrastive learning with cross-modal attention mechanisms to fully exploit both inter-modal and intra-modal relationships, as well as their complementary characteristics. As depicted in Figure 1, the proposed framework consists of three interconnected phases: pre-training, fine-tuning, and testing.

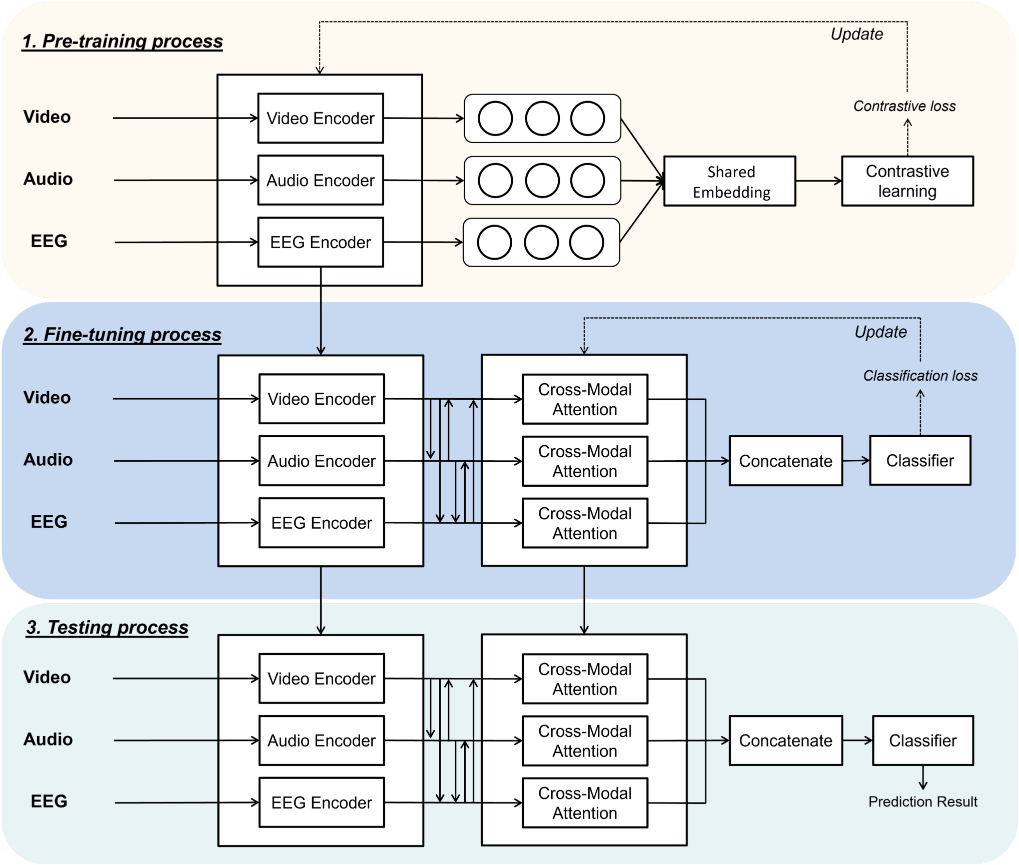

Figure 1. Diagram of a proposed multimodal framework which integrates video, audio, and EEG data for emotion recognition tasks. The framework consists of three main phases: pre-training, fine-tuning, and testing. In the pre-training phase, a modality encoder extracts features and fuses them into a combined embedding using contrastive learning. In the fine-tuning phase, cross-modal attention is utilized to capture interactions between modalities and trained for emotion recogntion. Finally, the test phase is used to obtain predictive results.

In the pre-training phase, modality-specific encoders are used to extract spatiotemporal features from the visual, audio, and EEG inputs. These features are then optimized in a supervised contrastive learning framework [55] to align the shared embedding spaces of the three modalities, allowing the model to learn more discriminative features for emotion recognition.

The fine-tuning phase incorporates the pre-trained encoders, cross-modal attention modules, and a task-specific classifier. Cross-modal attention is employed to capture and fuse salient feature information across modalities. The classifier is subsequently trained to predict emotion labels based on these multimodal fused representations.

During the testing phase, the fine-tuned encoders and cross-modal attention modules process the input data, and the classifier utilizes the fused multimodal representations to accurately predict emotional states.

3.1. Multimodal Encoder

This section describes the construction of a multimodal dataset and the encoder architecture used to extract spatiotemporal representations from visual, audio, and EEG data. The overall process involves two main steps: preprocessing and spatiotemporal encoding, as illustrated in Figure 2.

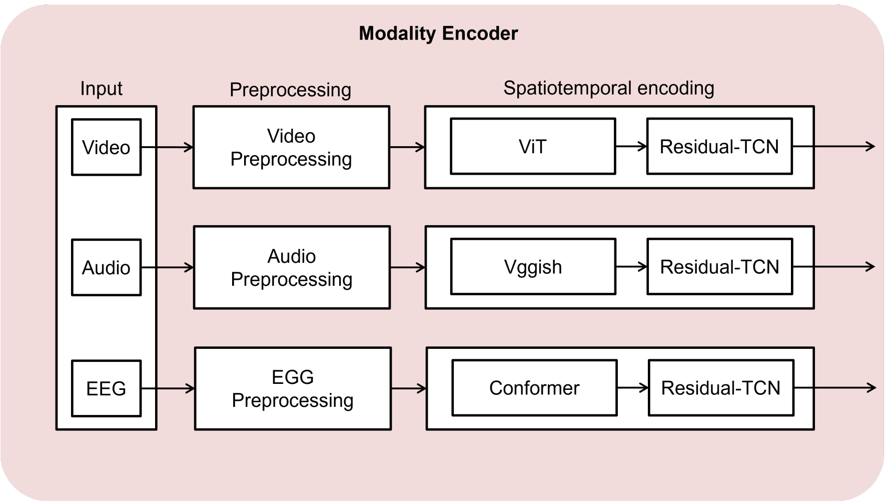

Figure 2. Illustration of modality-specific encoders for extracting spatiotemporal features from video, audio, and EEG data. Each modality is preprocessed before being input to the corresponding encoder: ViT for video, Vggish for audio, and Conformer for EEG. These encoded features are then processed using a Residual-TCN.

First, we recorded EEG signals while a subject watched videos from a database. The resulting dataset $X = \{(x_v, x_e, x_a, y)_i\}_{i=1}^{N}$ consisted of N samples, where x_v, x_e, and x_a represent visual, EEG, and audio data, respectively, and y is the corresponding label.

In the preprocessing stage, we segmented each video into 4 s blocks with 50% overlap. A demultiplexer separated the audio and visual signals, and we aligned the EEG segments with these blocks to ensure all modalities were synchronized and ready for feature extraction. To address the synchronization of these heterogeneous data types, we specifically applied a time-lagged synchronization approach [56]. This process ensured all modalities were temporally aligned and ready for feature extraction, minimizing any potential impact on the model's performance due to misalignment.

EEG signals underwent bandpass filtering to extract five distinct frequency bands: delta (0.4–4 Hz), theta (4–8 Hz), alpha (8–13 Hz), beta (13–30 Hz), and gamma (30–45 Hz). For each frequency band, the power spectral density (PSD) was calculated.

Audio signals were processed by applying a 50 ms Hamming window with a 10 ms hop length. This was followed by a short-time Fourier transform (STFT) on each windowed segment to extract the frequency components. The frequency components were then mapped to the Mel scale using 20 Mel filter banks to better align with human auditory perception. Finally, the Mel spectrogram was converted into a log-Mel spectrogram (LMS) by applying a logarithmic function.

For spatiotemporal encoding, we employed modality-specific encoder modules, as shown in Figure 2. Each encoder consisted of two main components: spatial encoding and temporal encoding. The spatial encoding used modality-specific models to capture spatial features, while the temporal encoding utilized a common residual temporal convolutional network (Residual-TCN) across all modalities.

3.1.1. Spatial Encoder

The spatial encoding process aims to extract meaningful spatial features from each modality, considering their unique characteristics. The input to the spatial encoding stage is the preprocessed data for each modality: video frames x_v, audio Mel spectrograms x_a, and the EEG power spectral density x_e.

For the visual data, we used a pre-trained vision transformer (ViT) [57], denoted as f_v, to extract spatial features from each video frame. The output of the visual spatial encoding, s_v, had a shape of $\mathbb{R}^{T_v \times D_v}$, where T_v is the number of video frames and D_v is the feature dimension of the visual spatial encoding.

Similarly, for the audio data, we applied a pre-trained Vggish [58] convolutional neural network, denoted as f_a, to extract features from the Mel spectrogram. The output of the audio spatial encoding, s_a, had a shape of $\mathbb{R}^{T_a \times D_a}$, where T_a is the number of audio frames and D_a is the feature dimension of the audio spatial encoding.

For the EEG data, we used a modified Conformer [59] neural network, denoted as f_e, to extract spatial features from the power spectral density, excluding the self-attention mechanism. The output of the EEG spatial encoding s_e had a shape of $\mathbb{R}^{T_e \times D_e}$, where T_e is the number of EEG segments and D_e is the feature dimension of the EEG spatial encoding.

The spatial encoding process maps the input data from each modality to a feature space which captures the essential spatial information. The encoded spatial features s_v, s_a, and s_e serve as the input to the subsequent temporal encoding stage.

3.1.2. Temporal Encoder

After spatial encoding, we employed a Residual-TCN for temporal encoding across all modalities. The Residual-TCN takes the spatially encoded features s_v, s_a, and s_e as input, extracting the temporal dependencies within each modality. The Residual-TCN effectively extracts and represents the temporal features from the video [60–62], audio [63,64], and EEG [65–67], enabling it to capture the unique temporal dynamics of each modality.

The network is composed of multiple blocks which process and refine temporal features. Each block begins with a dilated convolution layer, expanding the receptive field to capture long-range dependencies without increasing the parameter count. The dilated convolution is defined as

$$F(t) = \sum_{i=0}^{k-1} f(i) \cdot x(t - d \cdot i) \tag{1}$$

where $F(t)$ is the output at time t, f is the filter of a size k, x is the input, and d is the dilation rate. This operation allows the network to efficiently capture long-range dependencies by progressively increasing the dilation rate across layers. The dilation rate d determines the spacing between input elements in the convolution, enabling the network to expand its receptive field exponentially with depth while maintaining computational efficiency.

After the dilated convolution, a batch normalization layer stabilizes the learning process, followed by an ELU activation function. A dropout layer (rate of 0.5) prevents overfitting, and a residual connection facilitates smooth gradient flow. We stacked four such blocks, progressively increasing the dilation rate (1, 2, 4, and 8) to model both short- and long-term temporal patterns.

The final outputs of the Residual-TCN for visual, audio, and EEG modalities are represented by $h_v \in \mathbb{R}^{B \times T_v \times D}$, $h_a \in \mathbb{R}^{B \times T_a \times D}$, and $h_e \in \mathbb{R}^{B \times T_e \times D}$, where B is the batch size, T is the sequence length, and D is the feature dimension. These representations capture rich spatiotemporal information, forming a strong foundation for the subsequent contrastive learning phase, which enables the model to learn discriminative features for multimodal emotion recognition.

3.2. Contrastive Learning for Multimodal Representation

Contrastive learning has rapidly gained traction as an effective technique for aligning the embedding spaces of multimodal data [68–71]. This approach is particularly effective in multimodal contexts due to its ability to learn shared representations across different modalities, handle heterogeneous data, exploit cross-modal correspondences, and provide robustness to modality-specific noise. By learning discriminative representations, contrastive learning enhances the ability to distinguish between classes by pulling semantically similar samples closer in the latent space and pushing dissimilar samples apart.

To fully leverage contrastive learning in multimodal emotion recognition, it is essential to account for both intra-modal relationships within each modality and inter-modal interactions across modalities. Intra-modal learning captures the distinct characteristics inherent to each modality, ensuring that the model recognizes subtle patterns specific to visual, audio, or EEG data. Inter-modal learning, on the other hand, focuses on the complementarities between modalities, allowing the model to integrate and enhance shared information across different sources. By addressing both intra- and inter-modal relationships, as illustrated in Figure 3, the model becomes more capable of learning robust and discriminative representations, ultimately improving its ability to accurately classify emotions.

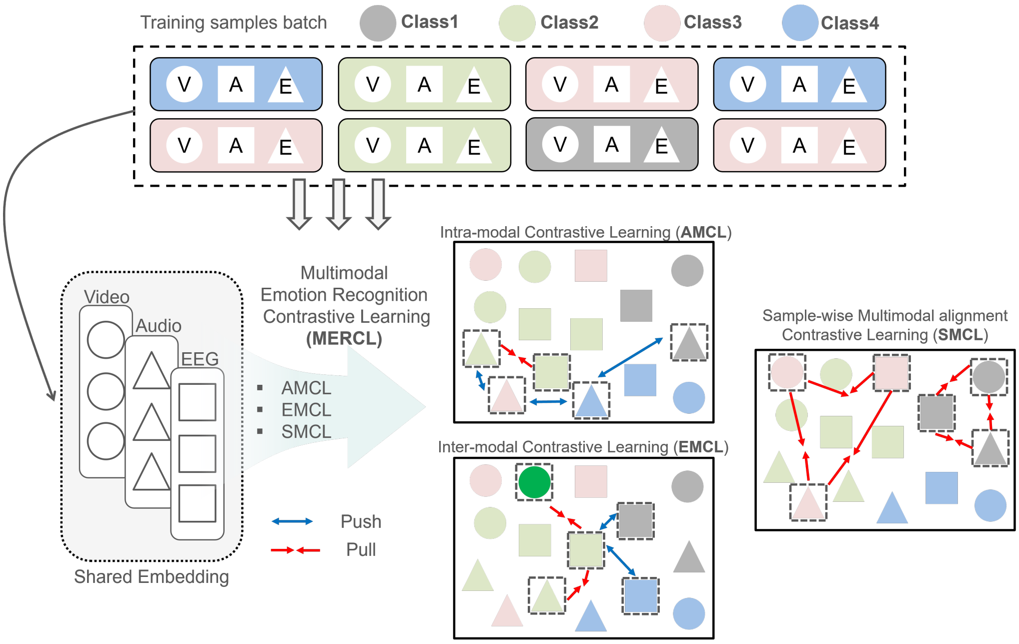
Training samples batch
Class1
Class2
Class3
Class4
V
A
E
Video
Audio
EEG
Shared Embedding
Multimodal Emotion Recognition Contrastive Learning (MERCL)
AMCL
EMCL
SMCL
Push
Pull
Intra-modal Contrastive Learning (AMCL)
Inter-modal Contrastive Learning (EMCL)
Sample-wise Multimodal alignment Contrastive Learning (SMCL)

Figure 3. MERCL consists of three components. (1) AMCL learns class-specific relationships within the same modality. (2) EMCL aligns representations across different modalities within the same sample. (3) Finally, SMCL minimizes modality gaps by aligning representations of different modalities within the same sample.

For this phase, we first projected the modality-specific spatiotemporal representations h_v, h_a, h_e obtained from the encoding module into a shared embedding space by passing them through a projection layer:

$$z_m = W_m h_m, \quad m \in \{v, a, e\} \tag{2}$$

where z_m represents the projected representation in the shared embedding space for the modality m. W_m denotes the projection matrix specific to each modality, and h_m is the original spatiotemporal representation for that modality.

(1) Intra-Modal Contrastive Learning (AMCL): AMCL focuses on learning class-specific relationships within a single modality by using supervised contrastive learning. For each minibatch, a set $C = \{p_1^m, p_2^m, \ldots, p_N^m, n_1^m, n_2^m, \ldots, n_M^m\}$ containing N positive samples and M negative samples is generated for a modality m. The AMCL loss is defined as

$$L_{AMCL} = -E_c\left[\log \frac{\sum_{i=1}^{N}(a^m)^T p_i^m}{\sum_{i=1}^{N}(a^m)^T p_i^m + \sum_{j=1}^{M}(a^m)^T n_j^m}\right], m \in \{v, a, e\} \tag{3}$$

where a^m is the anchor representation and p_i^m and n_j^m are the positive and negative samples, respectively, within the same modality m. This encourages representations of the same class to cluster closer together and those of different classes to be further apart within each modality.

(2) Inter-Modal Contrastive Learning (EMCL): EMCL focuses on learning relationships across different modalities by using supervised contrastive learning. Unlike AMCL, which defines positive and negative pairs within the same modality, EMCL defines these pairs across different modalities. For each minibatch, a set $C = \{p_1, p_2, \ldots, p_{2N}, n_1, n_2, \ldots, n_{2M}\}$

containing $2N$ positive samples and $2M$ negative samples is generated for an anchor modality m. The $2N$ positive and $2M$ negative samples are considered because EMCL explores interactions between the anchor modality and two other modalities, creating more pairings. The EMCL loss is defined as

$$L_{EMCL} = -E_c\left[\log \frac{\sum_{i=1}^{2N}(a^m)^T p_i}{\sum_{i=1}^{2N}(a^m)^T p_i + \sum_{j=1}^{2M}(a^m)^T n_j}\right], m \in \{v, a, e\} \tag{4}$$

where a^m is the anchor representation from the modality m and p_i and n_j are the positive and negative samples from other modalities, respectively. This loss encourages representations of the same class to be closer across different modalities while pushing apart representations of different classes.

However, while these two methods effectively capture intra- and inter-modal relationships, they may not fully address significant modality gaps within the same sample. Inspired by [72], we introduce an approach which minimizes this gap by aligning different modalities within the same sample, focusing solely on positive pairs. Negative pairs were excluded to prevent over-separation of modalities, which could lead to the loss of valuable modality-specific information. By aligning modalities with positive pairs, this approach preserves their unique characteristics, ensuring a balanced and informative representation. Therefore, we adopted this approach with a key modification. In our dataset, visual information is consistently available, but audio may be missing in cases like silent scenes in films. To address this, we calculated the energy of the audio signal to determine its presence and exclude samples with low or absent audio energy. This adjustment ensured that our method remains robust and adaptable to the specific characteristics of multimodal datasets.

(3) Sample-wise Multimodal Alignment Contrastive Learning (SMCL): SMCL focuses on minimizing the gap between representations of different modalities within the same sample. It only considers positive pairs, defined as embeddings from different modalities within the same sample. For each minibatch, a set $C = \{p_1^{m_2}, p_2^{m_3}\}$ is generated, where $m_1 \neq m_2 \neq m_3$ and $m_1, m_2, m_3 \in \{v, a, e\}$. SMCL also measures the energy of the audio signal and excludes low-energy samples to ensure robustness when the audio modality is missing or unreliable. The SMCL loss function is defined as

$$L_{SMCL} = E_c\left[\frac{1}{2}\sum_{i=1}^{2}\left|\left|(z^m)^T p_i^{m_i} - \alpha\right|\right|^2\right], \quad m \in \{v, a, e\} \tag{5}$$

$$[L_{SMCL} = E_c\left[\frac{1}{2}\left(\left|\left|(z^m)^T p_1^{m_1} - \alpha\right|\right|^2 + \left|\left|(z^m)^T p_2^{m_2} - \alpha\right|\right|^2\right)\right]] \tag{6}$$

where z^m is the anchor representation from the modality m and $p_i^{m_i}$ represents positive pairs from other modalities. A modality margin α accommodates minor differences between modalities while ensuring alignment. Minimizing this loss helps the model align representations across modalities within each sample, effectively reducing the modality gap. Here, i is considered to be up to two because SMCL aims to align pairs of modalities within the same sample. This reflects the two different modalities being aligned in each positive pair.

Together, these three loss functions are called multimodal emotion recognition contrastive learning (MERCL) loss and are defined as follows:

$$L_{MERCL} = \lambda_1 L_{AMCL} + \lambda_2 L_{EMCL} + \lambda_3 L_{SMCL} \tag{7}$$

where λ_1, λ_2, and λ_3 are hyperparameters that control the contribution of each loss function.

3.3. Cross-Modality Attention and Classifier

Cross-modality attention mechanisms can help two different modalities share the most important parts of each other, exploiting the complementarity between modalities to learn paired feature information and improve emotion recognition performance. In this paper, we employ a pairwise cross-modality attention [73–76] method to process hidden representations obtained through multimodal contrastive learning and identify deep connections between different modalities. This allows the model to provide a more holistic view of the information for emotion recognition, contributing to better performance.

As depicted in Figure 4, the cross-modal attentions are positioned after fixating the pre-trained encoder, which computes the cross-modality attention (CMA) for the spatiotemporal representations h_v, h_a, and h_e of each modality output by the encoder. The CMA is calculated using the multi-head attention (MHA) mechanism for each pair of modalities as follows:

- Video-Audio CMA:

$$\mathrm{CMA}^{a \to v} = \mathrm{MHA}(h_v, h_a), \quad \mathrm{CMA}^{v \to a} = \mathrm{MHA}(h_a, h_v) \tag{8}$$

- Video-EEG CMA:

$$\mathrm{CMA}^{e \to v} = \mathrm{MHA}(h_v, h_e), \quad \mathrm{CMA}^{v \to e} = \mathrm{MHA}(h_e, h_v) \tag{9}$$

- Audio-EEG CMA:

$$\mathrm{CMA}^{e \to a} = \mathrm{MHA}(h_a, h_e), \quad \mathrm{CMA}^{a \to e} = \mathrm{MHA}(h_e, h_a) \tag{10}$$

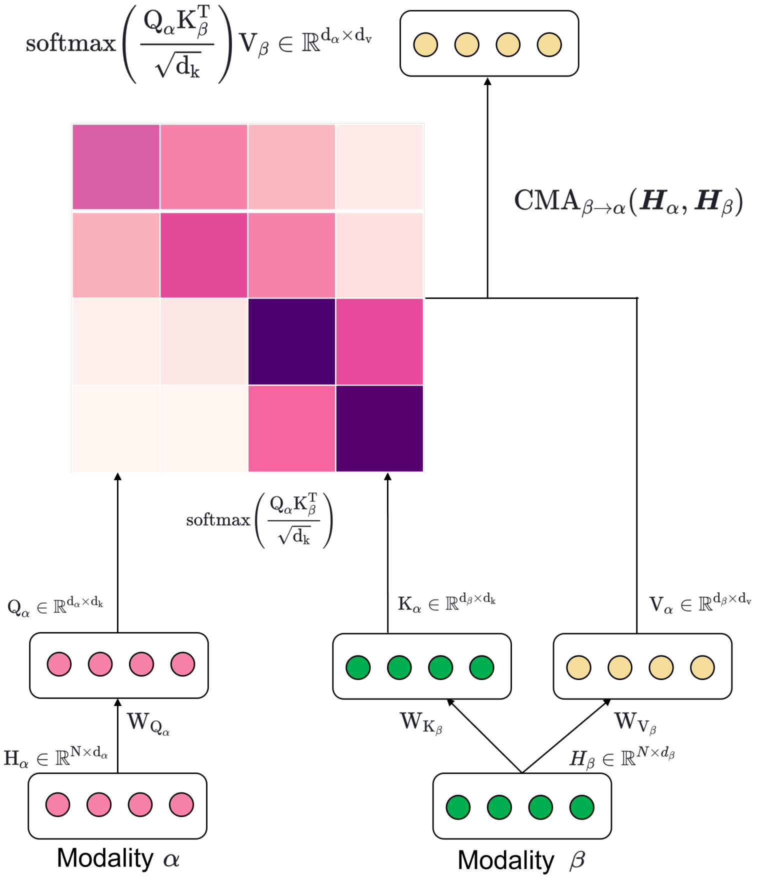
$\mathrm{softmax}\left(\frac{\mathrm{Q}_\alpha \mathrm{K}_\beta^{\mathrm{T}}}{\sqrt{\mathrm{d_k}}}\right)\mathrm{V}_\beta \in \mathbb{R}^{\mathrm{d}_\alpha \times \mathrm{d_v}}$
$\mathrm{CMA}_{\beta \to \alpha}(\boldsymbol{H}_\alpha, \boldsymbol{H}_\beta)$
$\mathrm{softmax}\left(\frac{\mathrm{Q}_\alpha \mathrm{K}_\beta^{\mathrm{T}}}{\sqrt{\mathrm{d_k}}}\right)$
$\mathrm{Q}_\alpha \in \mathbb{R}^{\mathrm{d}_\alpha \times \mathrm{d_k}}$
$\mathrm{K}_\alpha \in \mathbb{R}^{\mathrm{d}_\beta \times \mathrm{d_k}}$
$\mathrm{V}_\alpha \in \mathbb{R}^{\mathrm{d}_\beta \times \mathrm{d_v}}$
$\mathrm{W}_{\mathrm{Q}_\alpha}$
$\mathrm{W}_{\mathrm{K}_\beta}$
$\mathrm{W}_{\mathrm{V}_\beta}$
$\mathrm{H}_\alpha \in \mathbb{R}^{\mathrm{N} \times \mathrm{d}_\alpha}$
$H_\beta \in \mathbb{R}^{N \times d_\beta}$
Modality α
Modality β

Figure 4. Illustration of the CMA module between modalities α and β.

The MHA process allows the model to jointly attend to information from different representation subspaces at different positions. Each head in MHA can focus on different aspects of the input, enabling the model to capture various types of relationships between modalities. The MHA process for each pair of modalities can be described as follows, using $\text{CMA}^{a \to v}$ as an example:

$$Q_i = h_v W_{Q_i}^v, \quad K_i = h_a W_{K_i}^a, \quad V_i = h_a W_{V_i}^a \tag{11}$$

First, the input vectors h_v and h_a are linearly transformed into Query, Key, and Value vectors using learnable weight matrices $W_{Q_i}^v$, $W_{K_i}^a$, and $W_{V_i}^a$, respectively. These weight matrices play a crucial role in projecting the input features into different subspaces, allowing the model to capture various aspects of the inter-modal relationships:

$$A_i = \frac{Q_i K_i^T}{\sqrt{d_k}} V_i \tag{12}$$

The attention weights are then calculated by taking the dot product of the Query and Key vectors, scaling the result by the square root of the Key vector's dimension, and multiplying by the Value vector. This operation allows the model to determine the relevance of each part of the input from one modality to another, effectively capturing the cross-modal interactions:

$$\text{CMA}^{a \to v} = \text{Concatenation}(A_1, A_2, ..., A_M) W_O \tag{13}$$

The attention outputs from each head are concatenated and linearly transformed using a weight matrix W_O to obtain the final multi-head attention result $\text{CMA}^{a \to v}$. This combination of multiple attention heads allows the model to capture different types of cross-modal relationships simultaneously, enhancing its ability to understand complex inter-modal interactions. The same process is applied to compute the CMA for the other direction ($\text{CMA}^{v \to a}$) and for the other pairs of modalities (Video-EEG and Audio-EEG). This bidirectional attention mechanism ensures that the model captures the mutual influence between modalities, rather than just the influence of one modality on another.

The CMA outputs for each pair of modalities $(r_1, r_2, ..., r_6)$ are then concatenated into a unified vector $Output_{concatenation} = [r_1, r_2, ..., r_6]$, which contains the interaction information between all pairs of modalities.

This concatenated output represents a rich, multi-faceted representation of the cross-modal interactions, capturing both the individual modal characteristics and their inter-modal relationships.

Finally, the $Output_{concatenation}$ vector is fed into a multilayer perceptron (MLP) classifier, which consists of a linear layer followed by ReLU activation and a softmax function, to classify the emotion $\hat{y}$. The MLP classifier learns to interpret the complex cross-modal interactions captured by the CMA mechanism, mapping them to emotion categories. This final stage of the model effectively translates the intricate inter-modal relationships into meaningful emotion predictions.

The encoder module, cross-modal attention, and classifier are all optimized through the fine-tuning process using the same dataset used for pre-training. This end-to-end optimization ensures that all components of the model work together coherently to improve emotion recognition performance, leveraging the complementary information from different modalities.

Algorithm 1 presents a concise overview of our proposed multimodal emotion recognition method. This algorithm integrates the pre-training and fine-tuning stages, showcasing the key steps of our approach, including multimodal encoding, projection, contrastive learning, and cross-modal attention.

Algorithm 1 Multimodal emotion recognition training algorithm.

```
Require: Multimodal dataset $\mathcal{D} = \{(x_v, x_a, x_e, y)_i\}_{i=1}^{N}$
Ensure: Trained model parameters $\theta_{\text{encoder}}, \theta_{\text{CMA}}, \theta_{\text{classifier}}$
  // Stage 1: Pre-training encoders
  for each pre-training epoch do
    for each mini-batch $(x_v, x_a, x_e) \in \mathcal{D}$ do
      $(z_v, z_a, z_e) \leftarrow \text{Encode\_and\_Project}(x_v, x_a, x_e)$
      $\mathcal{L}_{\text{MERCL}} \leftarrow \lambda_1 \mathcal{L}_{\text{AMCL}} + \lambda_2 \mathcal{L}_{\text{EMCL}} + \lambda_3 \mathcal{L}_{\text{SMCL}}$
      $\theta_{\text{encoder}} \leftarrow \text{Update}(\theta_{\text{encoder}}, \nabla \mathcal{L}_{\text{MERCL}})$
    end for
  end for
  // Stage 2: Fine-tuning with CMA and classifier
  Freeze $\theta_{\text{encoder}}$
  for each fine-tuning epoch do
    for each mini-batch $(x_v, x_a, x_e, y) \in \mathcal{D}$ do
      $(h_v, h_a, h_e) \leftarrow \text{Encode}(x_v, x_a, x_e)$
      // Cross-Modality Attention between modality pairs
      $\text{CMA}^{a \to v} = \text{MHA}(h_v, h_a), \text{CMA}^{v \to a} = \text{MHA}(h_a, h_v)$
      $\text{CMA}^{e \to v} = \text{MHA}(h_v, h_e), \text{CMA}^{v \to e} = \text{MHA}(h_e, h_v)$
      $\text{CMA}^{e \to a} = \text{MHA}(h_a, h_e), \text{CMA}^{a \to e} = \text{MHA}(h_e, h_a)$
      Concatenate all CMA outputs:
      $Output_{\text{concatenation}} = [\text{CMA}^{a \to v}, \text{CMA}^{v \to a}, \text{CMA}^{e \to v}, \text{CMA}^{v \to e}, \text{CMA}^{e \to a}, \text{CMA}^{a \to e}]$
      $\hat{y} \leftarrow \text{Classifier}(\text{Concatenate}(Output_{\text{concatenation}}))$
      $\mathcal{L}_{\text{cls}} \leftarrow \text{Cross entropy}(\hat{y}, y)$
      $\theta_{\text{CMA}}, \theta_{\text{classifier}} \leftarrow \text{Update}(\theta_{\text{CMA}}, \theta_{\text{classifier}}, \nabla \mathcal{L}_{\text{cls}})$
    end for
  end for
  return $\theta_{\text{encoder}}, \theta_{\text{CMA}}, \theta_{\text{classifier}}$
```

4. Experimental Results

In this study, we evaluated the efficacy of our proposed method using four distinct datasets: DEAP [77], SEED [78], DEHBA, and MTIY. These datasets were selected for their comprehensive collection of EEG data elicited by audiovisual emotional stimuli.

4.1. Evaluation Datasets

- DEAP: The DEAP dataset contains EEG and peripheral signals collected from 32 participants (16 males and 16 females between the ages of 19 and 37). EEG signals were recorded while each participant watched 40 music video clips. Each participant rated their level of arousal, valence, dominance, and preference on a continuous scale from 1 to 9 using a Self-Assessment Manikin (SAM). Each trial contained 63 s of EEG signals, with the first 3 s serving as the baseline signal. The EEG signals were recorded at a sampling rate of 512 Hz using 32 electrodes. For this study, EEG data from 20 participants (10 males and 10 females) were selected for the experiment.
- SEED: The SEED dataset contains EEG and eye movement signals collected from 15 participants (7 males and 8 females). For this study, data from 10 participants (5 males and 5 females) were selected. Each participant's EEG signals were collected while watching 15 Chinese movie clips approximately 4 min in length, designed to evoke positive, neutral, and negative emotions. The signals collected from 62 electrodes had a sampling rate of 1 kHz, which was then downsampled to 200 Hz. After watching each film clip, each participant recorded an emotion label for each video as negative (-1), neutral (0), or positive (1).
- DEHBA: The DEHBA dataset is a human EEG dataset collected during emotional audiovisual stimulation. EEG data were measured while subjects watched video clips designed to elicit four emotional states: (1) happy, (2) sad, (3) angry, and (4) relaxed. These states are defined on a plane with axes representing arousal and valence from

the circumplex model of affect: "happy" corresponds to high valence and high arousal (HVHA), "angry" corresponds to low valence and high arousal (LVHA) "sad" corresponds to low valence and low arousal (LVLA), and "relaxed" corresponds to high valence and low arousal (HVLA).
Researchers selected 100 videos (25 for each emotional state) based on their ability to elicit strong emotions without relying on language understanding. These videos were validated by 30 college students, who rated the intensity of their emotions after viewing each clip. EEG data were collected from 30 participants using a 36 channel electrode cap at a sampling rate of 1 kHz, and for this study, data from 12 participants (6 males and 6 females) were selected for analysis.
The participants reported their emotional responses and rated the intensity of the emotions they experienced after viewing each video. This feedback was used to refine data selection and evaluate the results.

- MTIY: The Movie Trailer In YouTube (MTIY) dataset was constructed from 50 movie trailer videos retrieved from YouTube using the search term "movie trailer". The videos covered five genres—science fiction, comedy, action, horror, and romance—with 10 videos in each genre, and each video was 60 s long. Subjects were instructed to watch all 50 videos, and an Emotiv headset was used to obtain EEG signals, with EEG features extracted every second. The EEG data were collected using 14 electrodes. The EEG features were collected using 36 electrodes. For this study, data from 16 participants (8 males and 8 females) were used. Each subject rated the level of arousal, valence, dominance, and preference on a continuous scale from 1 to 9 after watching all of the videos.

4.2. Experimental Set-Up

To evaluate the emotion recognition performance of the proposed method and compare it with other existing approaches, we employed fourfold cross-validation to ensure robust evaluation. For comparison, we used existing multimodal models with both feature-level fusion and decision-level fusion methods. The following baseline methods were applied in the experiments:

- VE-BiLSTM [79]: This method employs a two-layer bidirectional LSTM network. It performs feature-level fusion by concatenating video and EEG features as input, where the video features are 1024 dimensional and the EEG features are also 1024 dimensional. The first LSTM layer has 1024 hidden units, and the second LSTM layer has 256 hidden units. The final recognition is performed using a softmax layer on top of the concatenated forward and backward hidden states from the second Bi-LSTM layer.
- AVE-RT [45]: This method combines EEG, audio, and visual features for emotion recognition through feature-level fusion. It extracts power spectral density features from the EEG signals across five frequency bands, audio features using eGeMAPS [80], and visual features, including the luminance coefficient and color energy. These multimodal features are concatenated at the feature level and fed into a random tree classifier for emotion recognition.
- AVE-KELM [81]: This method combines video content and EEG signals. It extracts audio-visual features from video clips and EEG features using wavelet packet decomposition (WPD). The video features are selected using double input symmetrical relevance (DISR), while EEG features are selected by a decision tree (DT). The selected features from both modalities are then combined at the decision level using a kernel-based extreme learning machine (ELM) for final emotion recognition.
- AVE-LSTM [82]: This method integrates the audio, video, and EEG modalities for emotion recognition. Each modality has its own feature extractor, and LSTM networks are used for emotion recognition. Specifically, audio features are derived from MFCC, video features are extracted using VGG19, and EEG features are obtained through PCA after bootstrapping. The outputs from each LSTM are individually used for emotion recognition, and the final emotion prediction is achieved through decision

fusion of these results. While the original approach also incorporated EMG data, our implementation excluded this modality.

All baseline methods were reimplemented according to the model configurations provided by their respective authors to ensure fair comparison. This approach allowed us to directly compare the performance of our proposed method with existing techniques under consistent experimental conditions.

The performance of each method was evaluated by using the accuracy and F1 score as the evaluation metrics:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{14}$$

$$\text{F1} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}} \tag{15}$$

where true positive (TP) is the number of positive samples correctly classified as positive, true negative (TN) is the number of negative samples correctly classified as negative, false positive (FP) is the number of negative samples incorrectly classified as positive, and false negative (FN) is the number of positive samples incorrectly classified as negative.

4.3. Experimental Results

This section provides a comprehensive analysis of the experimental results obtained from our proposed multimodal emotion recognition method. Our evaluation focuses on comparing the performance of our approach with that of a model combining EEG and audio-visual features, highlighting the effectiveness of our method. We explored how different modality combinations influence overall performance, revealing the impact of each on the model's efficacy. Through an ablation study, we demonstrate the significance of the contrastive learning and cross-modal attention mechanisms.

To evaluate our proposed model's performance, we conducted various experiments by using different emotion classification schemes for each dataset. Tables 1 and 2 illustrate our classification schemes for two-level and three-level emotions, respectively, as applied to the DEAP dataset. For the DEAP dataset, we redefined the emotion classes based on the original 1–9 rating scale for valence, arousal, and dominance. In addition to the two-level and three-level classifications shown in the tables, we implemented a four-level emotion classification by combining the valence and arousal dimensions. This resulted in four categories: HVHA, LVHA, LVLA, and HVLA. The SEED dataset, with its preexisting labels of negative (-1), neutral (0), and positive (1), was used for three-level valence classification experiments without any relabeling. For both the DEHBA and MTIY datasets, we employed the same four-level emotion classification scheme as that used with the DEAP dataset, categorizing emotions into HVHA, LVHA, LVLA, and HVLA based on the combination of valence and arousal dimensions.

Table 1. Emotion classes for two-level emotion classification on the DEAP dataset.

Rating Values (RVs)	Valence	Arousal	Dominance
$1 \leq \text{RVs} \leq 5$	Low	Low	Low
$6 \leq \text{RVs} \leq 9$	High	High	High

Table 2. Emotion classes for three-level emotion classification on the DEAP dataset.

Rating Values (RVs)	Valence	Arousal	Dominance
$1 \leq \text{RVs} \leq 3$	Negative	Activated	Controlled
$4 \leq \text{RVs} \leq 6$	Neutral	Moderate	Moderate
$7 \leq \text{RVs} \leq 9$	Positive	Deactivated	Overpowered

Tables 3–6 show the results of 2–4 levels of emotion classification using the four datasets. The final classification performance results were obtained by calculating the average of all the cross-validation folds.

Table 3. Performance comparison of different methods for two-level classification on the DEAP dataset.

Methods	Valence		Arousal		Dominance	
	Accuracy	F1	Accuracy	F1	Accuracy	F1
VE-BiLSTM [79]	71.8	71.4	70.1	70.2	71.5	71.3
AVE-KELM [81]	78.3	78.1	76.2	76.6	77.9	78.1
AVE-LSTM [82]	82.6	83.1	80.6	80.3	82.1	81.9
AVE-RT [45]	85.7	85.5	82.4	82.2	85.2	84.8
Proposed Method	93.4	93.2	91.7	92.0	93.5	93.2

Table 4. Performance comparison of different methods for three-level classification on the DEAP dataset.

Methods	Valence		Arousal		Dominance	
	Accuracy	F1	Accuracy	F1	Accuracy	F1
VE-BiLSTM [79]	64.6	64.2	64.3	63.9	63.7	63.5
AVE-KELM [81]	73.7	73.5	73.2	74.1	72.4	72.1
AVE-LSTM [82]	78.5	77.8	77.1	76.9	76.8	77.2
AVE-RT [45]	80.1	80.3	79.5	80.2	80.7	80.5
Proposed Method	89.3	89.6	88.6	88.2	89.2	89.5

Table 5. Performance comparison of different methods for four-level classification on the DEAP dataset and three-level classification on the SEED dataset.

Methods	DEAP: Four-Level		SEED: Three-Level	
	Accuracy	F1	Accuracy	F1
VE-BiLSTM [79]	60.1	59.4	69.3	70.2
AVE-KELM [81]	67.5	68.2	75.6	74.9
AVE-LSTM [82]	69.3	70.2	77.3	78.5
AVE-RT [45]	75.5	78.4	81.5	81.3
Proposed Method	83.2	84.1	90.9	91.2

Table 6. Performance comparison of different methods for four-level classification on the DEBHA dataset and MITY dataset.

Methods	DEBHA		MITY	
	Accuracy	F1	Accuracy	F1
VE-BiLSTM [79]	80.3	80.4	75.6	74.3
AVE-KELM [81]	83.4	82.7	78.3	79.2
AVE-LSTM [82]	85.3	84.1	80.2	81.1
AVE-RT [45]	87.5	86.6	82.6	83.0
Proposed Method	96.5	96.5	91.6	92.7

As shown in the results, the accuracy of emotion classification gradually decreased as the number of classes increased from two to four in the DEAP, SEED, DEBHA, and MITY datasets. This trend aligns with the expectation that as the number of emotion classes to be distinguished increases, the complexity of the patterns that the model needs to learn also increases.

The multimodal emotion recognition method proposed in this study, utilizing contrastive learning and cross-modal attention, consistently demonstrated superior performance compared with existing approaches. As can be seen in Tables 3–6, the proposed method achieved the highest accuracy and F1 scores across all datasets and classification levels. It should be noted that the higher performance in four-level classification for the DEBHA and MITY datasets compared with the DEAP and SEED datasets can be attributed to the selection of videos with more distinct emotional content in the DEBHA and MITY datasets. This clarity in emotional stimuli resulted in relatively consistent recognition performance across all methods except for our proposed approach, which showed significant improvement.

Existing feature-level fusion approaches (VE-BiLSTM and AVE-RT) simply concatenate features from multiple modalities, but this has limitations in fully capturing the complex interactions between modalities. This limitation becomes more apparent as the number of emotion classes increases. Another approach, decision-level fusion (e.g., AVE-KELM), processes each modality independently and combines them in the final decision stage. While this method can be computationally efficient, it has the drawback of potentially missing early interactions between modalities. In contrast, our proposed method can more effectively capture implicit correlations between modalities by explicitly learning the relationships between the features of each modality through contrastive learning.

Notably, the proposed method showed relatively less performance degradation as the number of classes increased. For example, when transitioning from two-level classification (Table 3) to four-level classification (Table 5) in the DEAP dataset, our method only showed about a 10% decrease in accuracy. In contrast, other methods showed a larger performance drop of 15–20%. This suggests that our approach is robust even in distinguishing more fine-grained emotional states.

Even compared with recent deep learning-based methods like AVE-LSTM, our method consistently showed better performance. This emphasizes that in multimodal emotion recognition, learning the unique characteristics and correlations between modalities is more important than simply using deep neural networks. Lastly, it is noteworthy that our method showed balanced performance improvement across all three emotional dimensions: valence, arousal, and dominance. This indicates that our approach can effectively capture the multidimensional nature of emotions.

We argue that combining external emotional stimuli (audiovisual data) with internal physiological responses (EEG signals) can provide a more comprehensive understanding of emotional states. To clearly understand this, we examined the effects of various modality combinations and investigated how audiovisual signals and EEG, which are in a stimulus–response relationship, interact to improve recognition accuracy.

Accordingly, in Table 7, the experimental results showed that emotion recognition based solely on audiovisual information has limitations in terms of accuracy and robustness. In contrast, when audio or video data were combined with EEG signals representing physiological responses, recognition performance significantly improved. This emphasizes the complementary role of external stimuli and internal physiological reactions in emotion recognition. Notably, among dual-modality combinations, the pairing of video data and EEG signals yielded the best performance, suggesting that visual cues provide particularly valuable information when combined with physiological data.

Table 7. Experiment with modality combinations on the DEBHA dataset.

Modality	Accuracy	F1
Audio + Video	78.4	77.8
Audio + EEG	82.5	81.9
Video + EEG	84.6	84.2
Audio + EEG + Video	96.5	96.5

The most notable point is that the highest accuracy in emotion recognition was achieved when all three modalities—audio, video, and EEG—were integrated. This result demonstrates the effectiveness of learning representations that incorporate both external emotional stimuli and internal physiological responses. These findings showcase the synergistic effect of combining multiple modalities, particularly the integration of external emotional stimuli (audiovisual data) with internal physiological responses (EEG signals). This approach provides a more comprehensive and accurate method for understanding and recognizing complex emotions.

While the modality combination experiments demonstrate the potential of a multimodal approach, they do not explain how the specific mechanisms of our proposed method maximize this potential. To explore this, we analyzed the contribution of each component through an ablation study. According to the experimental results in Table 8, performance dropped by 4.04% when contrastive learning was removed and by 2.19% when cross-modal attention was removed. When both elements were removed, the performance decreased the most (5.28%). This suggests that contrastive learning has a greater impact on learning rich information from multiple modalities compared with cross-modal attention.

Table 8. Ablation study of the impact of contrastive learning and cross-modal attention on the DEBHA dataset.

Condition	Accuracy	F1
Without Contrastive Learning	92.5	91.3
Without Cross-Modal Attention	94.3	94.0
Without Contrastive Learning and Cross-Modal Attention	91.2	92.1
Proposed Method	96.5	96.5

The importance of contrastive learning appears to stem from its ability to effectively capture complex relationships between different modalities and learn an integrated feature space. In particular, contrastive learning can be interpreted as playing a crucial role in learning subtle correlations between internal physiological responses like EEG signals and external expressions like audio-visual data.

The fact that performance dropped the most when both elements were removed shows that contrastive learning and cross-modal attention create a synergistic effect, maximizing the performance of multimodal emotion recognition. While contrastive learning learns the overall relationships between modalities, cross-modal attention enables more fine-grained information exchange based on this, thereby enhancing the model's expressiveness. These results demonstrate that our proposed method reaches beyond simply combining information from multiple modalities, effectively modeling complex interactions between each modality.

Finally, one important consideration in multimodal learning is how to handle samples that do not contain meaningful information in each modality. These samples can hinder model learning or lead to learning incorrect patterns. This issue becomes even more critical in complex tasks such as emotion recognition.

Table 9 shows our approach to this problem. When applying the audio energy-based sample selection method, the model's accuracy and F1 score improved. This proves that selectively using samples rich in information is more effective than simply using all samples. The key to this method is selecting samples likely to contain significant information based on the energy level of the audio signal. High-energy audio samples are generally more likely to contain clearer emotional expressions and are expected to have more distinct correlations with other modalities (EEG and video).

Table 9. Impact of audio energy-based sample selection on the DEBHA dataset.

Method	Accuracy	F1
Proposed method (all samples)	96.5	96.3
Proposed method (audio energy-based selection)	97.4	98.1

5. Discussion

While the proposed multimodal emotion recognition framework demonstrated significant improvements in classification accuracy, several important considerations and limitations warrant further discussion and future research.

The integration of multiple modalities and advanced techniques such as contrastive learning and cross-modal attention resulted in a highly complex model. This complexity, while contributing to the model's performance, poses challenges in terms of interpretability. Developing methods to visualize and explain the model's internal representations and decision boundaries could provide valuable insights and increase trust in the system's outputs. Furthermore, investigating how each modality's signals (video and audio) specifically influence brain responses as captured by EEG data is crucial. This exploration, along with existing efforts in interpreting multimodal emotion recognition systems [83–85], could provide insights into the actual interactions between different modalities and their impact on emotional responses. Such research could bridge the gap between computational models and neurophysiological processes, potentially leading to more biologically plausible and interpretable emotion recognition systems.

The current study utilized a limited number of datasets, which may affect the model's generalizability to diverse populations and contexts. The complex nature of the model, combined with limited data, raises concerns about potential overfitting. It is important to acknowledge the significant challenges in collecting comprehensive datasets for multimodal emotion recognition. Acquiring audiovisual materials which effectively elicit a wide range of emotions, along with corresponding EEG data, is a complex and resource-intensive process. The subjective nature of emotional responses and the variability across individuals further complicate this task. Future work should explore innovative approaches to data collection and augmentation, including semi-supervised learning techniques [86,87] which can leverage limited labeled data more effectively.

The proposed model's complexity necessitates substantial computational resources for training and inference, which may limit its applicability in real-time or resource-constrained environments. Future research should explore model compression techniques, such as knowledge distillation [88–90], to reduce the model's size and computational requirements without significantly compromising performance. Additionally, investigating incremental learning methods could facilitate more efficient model updates and adaptations to new data, enhancing the model's practical applicability in dynamic real-world scenarios.

6. Conclusions

This study proposed a novel multimodal approach for emotion recognition, integrating audio-visual data with EEG signals. Our research demonstrated that combining externally observable cues with internal physiological responses significantly improves emotion recognition accuracy. The proposed method, utilizing contrastive learning and cross-modal attention, consistently outperformed existing approaches across various datasets and classification levels.

Key findings include the crucial role of EEG signals in enhancing recognition accuracy, particularly when combined with audio-visual data, and the effectiveness of selective sample usage based on audio energy levels. Our approach showed robustness in distinguishing fine-grained emotional states, maintaining relatively high performance even as the number of emotion classes increased.

Future research should focus on enhancing model interpretability, personalizing emotion recognition. Developing techniques to visualize and explain the model's decision-making process will be crucial, particularly in understanding how different modalities interact and contribute to emotion classification. Investigating adaptive models which account for individual differences in emotional expression and perception could lead to more personalized and accurate systems. Additionally, exploring knowledge distillation methods to create simpler, more efficient models from our complex multimodal approach could address computational constraints while maintaining high performance. These advancements aim to create more interpretable, personalized, and efficient emotion recognition systems suitable for various real-world applications.

Author Contributions: Conceptualization, J.-H.L. and H.-G.K.; methodology, J.-H.L. and H.-G.K.; software, J.-H.L. and H.-G.K.; investigation, J.-H.L., J.-Y.K. and J.-H.L.; resources, J.-Y.K.; data curation, J.-H.L.; writing—original draft preparation, J.-H.L. and H.-G.K.; writing—review and editing, J.-Y.K.; visualization, J.-H.L.; project administration, H.-G.K.; funding acquisition, H.-G.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. NRF-2023R1A2C1006756).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The DEAP dataset can be found at https://www.eecs.qmul.ac.uk/mmv/datasets/deap/ (accessed on 5 July 2023). The SEED dataset is available at https://bcmi.sjtu.edu.cn/home/seed/ (accessed on 8 August 2023). Raw data from the DEHBA and MTIY datasets can be obtained by writing a formal email to Hyoung-Gook Kim.

Acknowledgments: This work was supported by the Research Grant of Kwangwoon University in 2024.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Andalibi, N.; Buss, J. The human in emotion recognition on social media: Attitudes, outcomes, risks. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 25–30 April 2020; pp. 1–16.
2. Dubey, A.; Shingala, B.; Panara, J.R.; Desai, K.; Sahana, M. Digital Content Recommendation System through Facial Emotion Recognition. *Int. J. Res. Appl. Sci. Eng. Technol* **2023**, *11*, 1272–1276. [CrossRef]
3. Pepa, L.; Spalazzi, L.; Capecci, M.; Ceravolo, M.G. Automatic emotion recognition in clinical scenario: A systematic review of methods. *IEEE Trans. Affect. Comput.* **2021**, *14*, 1675–1695. [CrossRef]
4. Caruelle, D.; Shams, P.; Gustafsson, A.; Lervik-Olsen, L. Affective computing in marketing: Practical implications and research opportunities afforded by emotionally intelligent machines. *Mark. Lett.* **2022**, *33*, 163–169. [CrossRef]
5. Jafari, M.; Shoeibi, A.; Khodatars, M.; Bagherzadeh, S.; Shalbaf, A.; García, D.L.; Gorriz, J.M.; Acharya, U.R. Emotion recognition in EEG signals using deep learning methods: A review. *Comput. Biol. Med.* **2023**, *165*, 107450. [CrossRef]
6. Lin, W.; Li, C. Review of studies on emotion recognition and judgment based on physiological signals. *Appl. Sci.* **2023**, *13*, 2573. [CrossRef]
7. Karnati, M.; Seal, A.; Bhattacharjee, D.; Yazidi, A.; Krejcar, O. Understanding deep learning techniques for recognition of human emotions using facial expressions: A comprehensive survey. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 1–31. [CrossRef]
8. Hashem, A.; Arif, M.; Alghamdi, M. Speech emotion recognition approaches: A systematic review. *Speech Commun.* **2023**, *154*, 102974. [CrossRef]
9. Mittal, T.; Mathur, P.; Bera, A.; Manocha, D. Affect2mm: Affective analysis of multimedia content using emotion causality. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 5661–5671.
10. Srivastava, D.; Singh, A.K.; Tapaswi, M. How You Feelin'? Learning Emotions and Mental States in Movie Scenes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 2517–2528.
11. Wang, S.; Ji, Q. Video affective content analysis: A survey of state-of-the-art methods. *IEEE Trans. Affect. Comput.* **2015**, *6*, 410–430. [CrossRef]
12. Wang, Y.; Song, W.; Tao, W.; Liotta, A.; Yang, D.; Li, X.; Gao, S.; Sun, Y.; Ge, W.; Zhang, W.; et al. A systematic review on affective computing: Emotion models, databases, and recent advances. *Inf. Fusion* **2022**, *83*, 19–52. [CrossRef]
13. Goncalves, L.; Busso, C. Robust audiovisual emotion recognition: Aligning modalities, capturing temporal information, and handling missing features. *IEEE Trans. Affect. Comput.* **2022**, *13*, 2156–2170. [CrossRef]
14. Ezzameli, K.; Mahersia, H. Emotion recognition from unimodal to multimodal analysis: A review. *Inf. Fusion* **2023**, *99*, 101847. [CrossRef]
15. Ahmed, N.; Al Aghbari, Z.; Girija, S. A systematic survey on multimodal emotion recognition using learning algorithms. *Intell. Syst. Appl.* **2023**, *17*, 200171. [CrossRef]
16. Wei, Y.; Hu, D.; Tian, Y.; Li, X. Learning in audio-visual context: A review, analysis, and new perspective. *arXiv* **2022**, arXiv:2208.09579.

17. Huang, Y.; Du, C.; Xue, Z.; Chen, X.; Zhao, H.; Huang, L. What makes multi-modal learning better than single (provably). *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 10944–10956.
18. Ma, Y.; Hao, Y.; Chen, M.; Chen, J.; Lu, P.; Košir, A. Audio-visual emotion fusion (AVEF): A deep efficient weighted approach. *Inf. Fusion* **2019**, *46*, 184–192. [CrossRef]
19. Hossain, M.S.; Muhammad, G. Emotion recognition using deep learning approach from audio-visual emotional big data. *Inf. Fusion* **2019**, *49*, 69–78. [CrossRef]
20. Ghaleb, E.; Popa, M.; Asteriadis, S. Metric learning-based multimodal audio-visual emotion recognition. *IEEE Multimed.* **2019**, *27*, 37–48. [CrossRef]
21. Praveen, R.G.; Granger, E.; Cardinal, P. Cross attentional audio-visual fusion for dimensional emotion recognition. In Proceedings of the 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021), Jodhpur, India, 15–18 December 2021; pp. 1–8.
22. Chen, S.; Tang, J.; Zhu, L.; Kong, W. A multi-stage dynamical fusion network for multimodal emotion recognition. *Cogn. Neurodyn.* **2023**, *17*, 671–680. [CrossRef]
23. Zali-Vargahan, B.; Charmin, A.; Kalbkhani, H.; Barghandan, S. Semisupervised Deep Features of Time-Frequency Maps for Multimodal Emotion Recognition. *Int. J. Intell. Syst.* **2023**, *2023*, 3608115. [CrossRef]
24. Perez-Gaspar, L.A.; Caballero-Morales, S.O.; Trujillo-Romero, F. Multimodal emotion recognition with evolutionary computation for human-robot interaction. *Expert Syst. Appl.* **2016**, *66*, 42–61. [CrossRef]
25. Kim, D.H.; Baddar, W.J.; Jang, J.; Ro, Y.M. Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition. *IEEE Trans. Affect. Comput.* **2017**, *10*, 223–236. [CrossRef]
26. Hao, M.; Cao, W.H.; Liu, Z.T.; Wu, M.; Xiao, P. Visual-audio emotion recognition based on multi-task and ensemble learning with multiple features. *Neurocomputing* **2020**, *391*, 42–51. [CrossRef]
27. Farhoudi, Z.; Setayeshi, S. Fusion of deep learning features with mixture of brain emotional learning for audio-visual emotion recognition. *Speech Commun.* **2021**, *127*, 92–103. [CrossRef]
28. Majumder, N.; Hazarika, D.; Gelbukh, A.; Cambria, E.; Poria, S. Multimodal sentiment analysis using hierarchical fusion with context modeling. *Knowl.-Based Syst.* **2018**, *161*, 124–133. [CrossRef]
29. Sarvestani, R.R.; Boostani, R. FF-SKPCCA: Kernel probabilistic canonical correlation analysis. *Appl. Intell.* **2017**, *46*, 438–454. [CrossRef]
30. Deldari, S.; Xue, H.; Saeed, A.; He, J.; Smith, D.V.; Salim, F.D. Beyond just vision: A review on self-supervised representation learning on multimodal and temporal data. *arXiv* **2022**, arXiv:2206.02353.
31. Vempati, R.; Sharma, L.D. A systematic review on automated human emotion recognition using electroencephalogram signals and artificial intelligence. *Results Eng.* **2023**, *18*, 101027. [CrossRef]
32. Rainville, P.; Bechara, A.; Naqvi, N.; Damasio, A.R. Basic emotions are associated with distinct patterns of cardiorespiratory activity. *Int. J. Psychophysiol.* **2006**, *61*, 5–18. [CrossRef]
33. Kreibig, S.D. Autonomic nervous system activity in emotion: A review. *Biol. Psychol.* **2010**, *84*, 394–421. [CrossRef]
34. Sarvakar, K.; Senkamalavalli, R.; Raghavendra, S.; Kumar, J.S.; Manjunath, R.; Jaiswal, S. Facial emotion recognition using convolutional neural networks. *Mater. Today Proc.* **2023**, *80*, 3560–3564. [CrossRef]
35. Ye, J.; Wen, X.C.; Wei, Y.; Xu, Y.; Liu, K.; Shan, H. Temporal modeling matters: A novel temporal emotional modeling approach for speech emotion recognition. In Proceedings of the ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; pp. 1–5.
36. Can, Y.S.; Mahesh, B.; André, E. Approaches, applications, and challenges in physiological emotion recognition—A tutorial overview. *Proc. IEEE* **2023**, *111*, 1287–1313. [CrossRef]
37. Chakravarthi, B.; Ng, S.C.; Ezilarasan, M.; Leung, M.F. EEG-based emotion recognition using hybrid CNN and LSTM classification. *Front. Comput. Neurosci.* **2022**, *16*, 1019776. [CrossRef] [PubMed]
38. Antoniadis, P.; Pikoulis, I.; Filntisis, P.P.; Maragos, P. An audiovisual and contextual approach for categorical and continuous emotion recognition in-the-wild. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–11 October 2021; pp. 3645–3651.
39. Zhang, Y.H.; Huang, R.; Zeng, J.; Shan, S. M 3 f: Multi-modal continuous valence-arousal estimation in the wild. In Proceedings of the 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), Buenos Aires, Argentina, 16–20 November 2020; pp. 632–636.
40. Mocanu, B.; Tapu, R.; Zaharia, T. Multimodal emotion recognition using cross modal audio-video fusion with attention and deep metric learning. *Image Vis. Comput.* **2023**, *133*, 104676. [CrossRef]
41. Udahemuka, G.; Djouani, K.; Kurien, A.M. Multimodal Emotion Recognition Using Visual, Vocal and Physiological Signals: A Review. *Appl. Sci.* **2024**, *14*, 8071. [CrossRef]
42. Li, Z.; Zhang, G.; Dang, J.; Wang, L.; Wei, J. Multi-modal emotion recognition based on deep learning of EEG and audio signals. In Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN), Virtual, 18–22 July 2021; pp. 1–6.
43. Song, B.C.; Kim, D.H. Hidden emotion detection using multi-modal signals. In Proceedings of the Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems, Yokohama, Japan, 8–13 May 2021; pp. 1–7.
44. Liang, Z.; Zhang, X.; Zhou, R.; Zhang, L.; Li, L.; Huang, G.; Zhang, Z. Cross-individual affective detection using EEG signals with audio-visual embedding. *Neurocomputing* **2022**, *510*, 107–121. [CrossRef]

45. Xing, B.; Zhang, H.; Zhang, K.; Zhang, L.; Wu, X.; Shi, X.; Yu, S.; Zhang, S. Exploiting EEG signals and audiovisual feature fusion for video emotion recognition. *IEEE Access* **2019**, *7*, 59844–59861. [CrossRef]
46. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 13–18 July 2020; pp. 1597–1607.
47. Akbari, H.; Yuan, L.; Qian, R.; Chuang, W.H.; Chang, S.F.; Cui, Y.; Gong, B. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 24206–24221.
48. Dissanayake, V.; Seneviratne, S.; Rana, R.; Wen, E.; Kaluarachchi, T.; Nanayakkara, S. Sigrep: Toward robust wearable emotion recognition with contrastive representation learning. *IEEE Access* **2022**, *10*, 18105–18120. [CrossRef]
49. Jiang, W.B.; Li, Z.; Zheng, W.L.; Lu, B.L. Functional emotion transformer for EEG-assisted cross-modal emotion recognition. In Proceedings of the ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 14–19 April 2024; pp. 1841–1845.
50. Tang, J.; Ma, Z.; Gan, K.; Zhang, J.; Yin, Z. Hierarchical multimodal-fusion of physiological signals for emotion recognition with scenario adaption and contrastive alignment. *Inf. Fusion* **2024**, *103*, 102129. [CrossRef]
51. Yang, D.; Huang, S.; Liu, Y.; Zhang, L. Contextual and cross-modal interaction for multi-modal speech emotion recognition. *IEEE Signal Process. Lett.* **2022**, *29*, 2093–2097. [CrossRef]
52. Praveen, R.G.; de Melo, W.C.; Ullah, N.; Aslam, H.; Zeeshan, O.; Denorme, T.; Pedersoli, M.; Koerich, A.L.; Bacon, S.; Cardinal, P.; et al. A joint cross-attention model for audio-visual fusion in dimensional emotion recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 2486–2495.
53. Zhao, J.; Ru, G.; Yu, Y.; Wu, Y.; Li, D.; Li, W. Multimodal music emotion recognition with hierarchical cross-modal attention network. In Proceedings of the 2022 IEEE International Conference on Multimedia and Expo (ICME), Taipei, Taiwan, 18–22 July 2022; pp. 1–6.
54. Praveen, R.G.; Alam, J. Recursive Joint Cross-Modal Attention for Multimodal Fusion in Dimensional Emotion Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 4803–4813.
55. Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; Krishnan, D. Supervised contrastive learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 18661–18673.
56. Xiao, R.; Ding, C.; Hu, X. Time Synchronization of Multimodal Physiological Signals through Alignment of Common Signal Types and Its Technical Considerations in Digital Health. *J. Imaging* **2022**, *8*, 120. [CrossRef] [PubMed]
57. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
58. Hershey, S.; Chaudhuri, S.; Ellis, D.P.; Gemmeke, J.F.; Jansen, A.; Moore, R.C.; Plakal, M.; Platt, D.; Saurous, R.A.; Seybold, B.; et al. CNN architectures for large-scale audio classification. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 131–135.
59. Gulati, A.; Qin, J.; Chiu, C.C.; Parmar, N.; Zhang, Y.; Yu, J.; Han, W.; Wang, S.; Zhang, Z.; Wu, Y.; et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv* **2020**, arXiv:2005.08100.
60. Shao, W.; Xiao, R.; Rajapaksha, P.; Wang, M.; Crespi, N.; Luo, Z.; Minerva, R. Video anomaly detection with NTCN-ML: A novel TCN for multi-instance learning. *Pattern Recognit.* **2023**, *143*, 109765. [CrossRef]
61. Singhania, D.; Rahaman, R.; Yao, A. C2F-TCN: A framework for semi-and fully-supervised temporal action segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 11484–11501. [CrossRef]
62. Zhou, W.; Lu, J.; Xiong, Z.; Wang, W. Leveraging TCN and Transformer for effective visual-audio fusion in continuous emotion recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 5756–5763.
63. Ishaq, M.; Khan, M.; Kwon, S. TC-Net: A Modest & Lightweight Emotion Recognition System Using Temporal Convolution Network. *Comput. Syst. Sci. Eng.* **2023**, *46*, 3355–3369.
64. Lemaire, Q.; Holzapfel, A. Temporal convolutional networks for speech and music detection in radio broadcast. In Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR 2019), Delft, The Netherlands, 4–8 November 2019.
65. Li, C.; Chen, B.; Zhao, Z.; Cummins, N.; Schuller, B.W. Hierarchical attention-based temporal convolutional networks for eeg-based emotion recognition. In Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 1240–1244.
66. Bi, J.; Wang, F.; Ping, J.; Qu, G.; Hu, F.; Li, H.; Han, S. FBN-TCN: Temporal convolutional neural network based on spatial domain fusion brain networks for affective brain–computer interfaces. *Biomed. Signal Process. Control* **2024**, *94*, 106323. [CrossRef]
67. Yang, L.; Wang, Y.; Ouyang, R.; Niu, X.; Yang, X.; Zheng, C. Electroencephalogram-based emotion recognition using factorization temporal separable convolution network. *Eng. Appl. Artif. Intell.* **2024**, *133*, 108011. [CrossRef]
68. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 18–24 July 2021; pp. 8748–8763.
69. Wang, Z.; Wu, Z.; Agarwal, D.; Sun, J. Medclip: Contrastive learning from unpaired medical images and text. *arXiv* **2022**, arXiv:2210.10163.

70. Guzhov, A.; Raue, F.; Hees, J.; Dengel, A. Audioclip: Extending clip to image, text and audio. In Proceedings of the ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual, 7–13 May 2022; pp. 976–980.
71. Geng, X.; Liu, H.; Lee, L.; Schuurmans, D.; Levine, S.; Abbeel, P. Multimodal masked autoencoders learn transferable representations. *arXiv* **2022**, arXiv:2205.14204.
72. Mai, S.; Zeng, Y.; Zheng, S.; Hu, H. Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. *IEEE Trans. Affect. Comput.* **2022**, *14*, 2276–2289. [CrossRef]
73. Huang, G.; Ma, F. Concad: Contrastive learning-based cross attention for sleep apnea detection. In *Proceedings of the Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track: European Conference, ECML PKDD 2021, Bilbao, Spain, 13–17 September 2021*; Proceedings, Part V 21; Springer: Berlin/Heidelberg, Germany, 2021; pp. 68–84.
74. Zhou, R.; Zhou, H.; Shen, L.; Chen, B.Y.; Zhang, Y.; He, L. Integrating Multimodal Contrastive Learning and Cross-Modal Attention for Alzheimer's Disease Prediction in Brain Imaging Genetics. In Proceedings of the 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Istanbul, Turkiye, 5–8 December 2023; pp. 1806–1811.
75. Nguyen, C.V.T.; Mai, A.T.; Le, T.S.; Kieu, H.D.; Le, D.T. Conversation Understanding using Relational Temporal Graph Neural Networks with Auxiliary Cross-Modality Interaction. *arXiv* **2023**, arXiv:2311.04507.
76. Krishna, D.; Patil, A. Multimodal Emotion Recognition Using Cross-Modal Attention and 1D Convolutional Neural Networks. In Proceedings of the Interspeech, Shanghai, China, 25–29 October 2020; pp. 4243–4247.
77. Koelstra, S.; Muhl, C.; Soleymani, M.; Lee, J.S.; Yazdani, A.; Ebrahimi, T.; Pun, T.; Nijholt, A.; Patras, I. Deap: A database for emotion analysis; using physiological signals. *IEEE Trans. Affect. Comput.* **2011**, *3*, 18–31. [CrossRef]
78. Zheng, W.L.; Lu, B.L. Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. *IEEE Trans. Auton. Ment. Dev.* **2015**, *7*, 162–175. [CrossRef]
79. Ogawa, T.; Sasaka, Y.; Maeda, K.; Haseyama, M. Favorite video classification based on multimodal bidirectional LSTM. *IEEE Access* **2018**, *6*, 61401–61409. [CrossRef]
80. Eyben, F.; Scherer, K.R.; Schuller, B.W.; Sundberg, J.; André, E.; Busso, C.; Devillers, L.Y.; Epps, J.; Laukka, P.; Narayanan, S.S.; et al. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Trans. Affect. Comput.* **2015**, *7*, 190–202. [CrossRef]
81. Duan, L.; Ge, H.; Yang, Z.; Chen, J. Multimodal fusion using kernel-based ELM for video emotion recognition. In *Proceedings of the ELM-2015 Volume 1: Theory, Algorithms and Applications (I)*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 371–381.
82. Chen, J.; Ro, T.; Zhu, Z. Emotion recognition with audio, video, EEG, and EMG: A dataset and baseline approaches. *IEEE Access* **2022**, *10*, 13229–13242. [CrossRef]
83. Asokan, A.R.; Kumar, N.; Ragam, A.V.; Shylaja, S. Interpretability for multimodal emotion recognition using concept activation vectors. In Proceedings of the 2022 International Joint Conference on Neural Networks (IJCNN), Padua, Italy, 18–23 July 2022; pp. 1–8.
84. Polo, E.M.; Mollura, M.; Lenatti, M.; Zanet, M.; Paglialonga, A.; Barbieri, R. Emotion recognition from multimodal physiological measurements based on an interpretable feature selection method. In Proceedings of the 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Virtual, 1–5 November 2021; pp. 989–992.
85. Liu, B.; Guo, J.; Chen, C.P.; Wu, X.; Zhang, T. Fine-grained interpretability for EEG emotion recognition: Concat-aided grad-CAM and systematic brain functional network. *IEEE Trans. Affect. Comput.* **2023**, *15*, 671–684. [CrossRef]
86. Zhao, S.; Hong, X.; Yang, J.; Zhao, Y.; Ding, G. Toward Label-Efficient Emotion and Sentiment Analysis. *Proc. IEEE* **2023**, *111*, 1159–1197. [CrossRef]
87. Qiu, S.; Chen, Y.; Yang, Y.; Wang, P.; Wang, Z.; Zhao, H.; Kang, Y.; Nie, R. A review on semi-supervised learning for EEG-based emotion recognition. *Inf. Fusion* **2023**, *104*, 102190. [CrossRef]
88. Ma, H.; Wang, J.; Lin, H.; Zhang, B.; Zhang, Y.; Xu, B. A transformer-based model with self-distillation for multimodal emotion recognition in conversations. *IEEE Trans. Multimed.* **2023**, *26*, 776–788. [CrossRef]
89. Aslam, M.H.; Pedersoli, M.; Koerich, A.L.; Granger, E. Multi Teacher Privileged Knowledge Distillation for Multimodal Expression Recognition. *arXiv* **2024**, arXiv:2408.09035.
90. Sun, T.; Wei, Y.; Ni, J.; Liu, Z.; Song, X.; Wang, Y.; Nie, L. Muti-modal Emotion Recognition via Hierarchical Knowledge Distillation. *IEEE Trans. Multimed.* **2024**, *26*, 9036–9046. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.