Multimodal Technologies and Interaction

Article

Cross-Modal Attention Fusion: A Deep Learning and Affective Computing Model for Emotion Recognition

Himanshu Kumar [1], Martin Aruldoss [1] and Martin Wynn [2,*]

[1] School of Mathematics and Computer Sciences, Central University of Tamil Nadu, Thiruvarur 610005, India; himanshukphd20@students.cutn.ac.in (H.K.); martin@cutn.ac.in (M.A.)
[2] School of Business, Computing and Social Sciences, University of Gloucestershire, Cheltenham GL50 2RH, UK
* Correspondence: mwynn@glos.ac.uk

Abstract

Artificial emotional intelligence is a sub-domain of human–computer interaction research that aims to develop deep learning models capable of detecting and interpreting human emotional states through various modalities. A major challenge in this domain is identifying meaningful correlations between heterogeneous modalities—for example, between audio and visual data—due to their distinct temporal and spatial properties. Traditional fusion techniques used in multimodal learning to combine data from different sources often fail to adequately capture meaningful and less computational cross-modal interactions, and struggle to adapt to varying modality reliability. Following a review of the relevant literature, this study adopts an experimental research method to develop and evaluate a mathematical cross-modal fusion model, thereby addressing a gap in the extant research literature. The framework uses the Tucker tensor decomposition to analyse the multi-dimensional array of data into a set of matrices to support the integration of temporal features from audio and spatiotemporal features from visual modalities. A cross-attention mechanism is incorporated to enhance cross-modal interaction, enabling each modality to attend to the relevant information from the other. The efficacy of the model is rigorously evaluated on three publicly available datasets and the results conclusively demonstrate that the proposed fusion technique outperforms conventional fusion methods and several more recent approaches. The findings break new ground in this field of study and will be of interest to researchers and developers in artificial emotional intelligence.

Keywords: artificial emotional intelligence; human–computer interaction; cross-attention mechanism; categorical emotions; spatiotemporal features; Tucker decomposition; cross-modal framework

Academic Editor: Stephan Schlögl

Received: 11 October 2025
Revised: 18 November 2025
Accepted: 21 November 2025
Published: 24 November 2025

Citation: Kumar, H.; Aruldoss, M.; Wynn, M. Cross-Modal Attention Fusion: A Deep Learning and Affective Computing Model for Emotion Recognition. *Multimodal Technol. Interact.* **2025**, *9*, 116. https://doi.org/10.3390/mti9120116

1. Introduction

Artificial emotion recognition is a subfield of affective computing that empowers machines to perceive, interpret, and respond to human emotions. Ekman's Theory of Basic Emotions identifies six universal emotions: anger, fear, sadness, disgust, happiness, and surprise [1]. As an essential component of human–computer interaction, emotion recognition plays a pivotal role in diverse applications, such as teaching and learning [2], online gaming [3], medical diagnostics [4], and decision-making assistance [5]. Conventional emotion recognition techniques have relied on handcrafted feature extraction processes applied to unimodal inputs, such as speech signals [6] or facial expressions [7]. Standard features such as pitch [8], energy, spectral properties [9], facial geometric features, facial landmarks [10], expression descriptors, and facial action coding units [11] have achieved some successes, but in general have often failed to represent the complexity and variability in human emotions in real-world scenarios.

The advent of deep learning has transformed this field of study by enabling models to extract hierarchical and discriminative features from raw data automatically. Popular deep learning architectures such as Convolutional Neural Networks (CNNs) [12], Recurrent Neural Networks (RNNs) [13], Long Short-Term Memory (LSTM) [14], and Bidirectional LSTM (Bi-LSTM) [15] have demonstrated strong capabilities in capturing spatial and temporal dynamics [16]. In particular, cross-modal emotion recognition, which integrates information from heterogeneous modalities, has emerged as a promising direction due to the complementary nature of these signals.

Cross-modal emotion recognition typically involves two core components: feature extraction [17] and feature fusion [18]. Feature extraction aims to derive salient and high-level representations from raw inputs, while fusion seeks to integrate these heterogeneous features into a unified representation. While traditional fusion strategies, such as early fusion (feature-level concatenation) [19], late fusion (decision-level concatenation) [20], and hybrid fusion [21] have achieved reasonable success, they are fundamentally limited in that they often rely on shallow architectures and struggle with noise or incomplete data from any one modality. Feature fusion preserves the information from each modality to a great extent, but it often introduces challenges such as temporal misalignment between modalities and a heightened risk of overfitting, due to the high dimensionality of the combined feature space.

In this context, this paper addresses the following research questions (RQs):

RQ1. To what extent do affective computing and recent advancements in deep learning improve cross-modal fusion for audio and image-based emotion recognition?

RQ2. Can a new model be developed and validated to improve correlations between heterogeneous modalities (such as audio and visual data) to detect and interpret human emotional states?

Following this brief introduction, Section 2 outlines the research methodology and more specifically describes the experimental design, deep-learning model architecture, and hyperparameter configurations. Section 3 sets out the main results of this study, whilst emergent issues are analysed and discussed in Section 4. Finally, Section 5 summarises the contribution of this research, points out limitations, and notes possible future research initiatives in this field of study.

2. Materials and Methods

This study comprises two main phases which combine different research methods and philosophies (Figure 1). In Phase 1, an interpretivist philosophy is assumed, combined with an inductive qualitative approach to assess the extant literature and develop a provisional conceptual framework as context for the subsequent primary research phase. Gill and Johnson [22] suggest that such an approach is most appropriate when the research aim is exploratory in nature, as was the case in Phase 1 of the project when the objective was to explore the relevant literature on affective computing. Then, in Phase 2, a positivist stance is adopted for the conduct and evaluation of the experiment. More detail on these two phases is provided below.

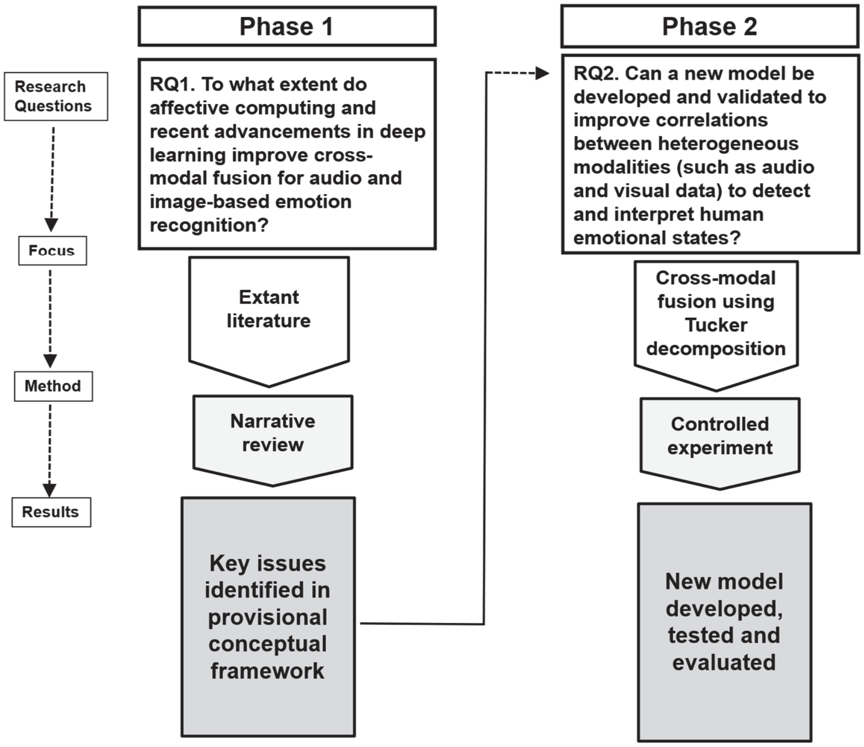
Phase 1
Phase 2
Research Questions
Focus
Method
Results
RQ1. To what extent do affective computing and recent advancements in deep learning improve cross-modal fusion for audio and image-based emotion recognition?
RQ2. Can a new model be developed and validated to improve correlations between heterogeneous modalities (such as audio and visual data) to detect and interpret human emotional states?
Extant literature
Cross-modal fusion using Tucker decomposition
Narrative review
Controlled experiment
Key issues identified in provisional conceptual framework
New model developed, tested and evaluated

Figure 1. The two-phase research process.

2.1. Narrative Review

Phase 1 comprised a narrative review of the relevant literature to identify key themes and provide the basis for development of a provisional conceptual framework (PCF) for the conduct of the experimental phase. Phase 1 was a “broad scan of contextual literature” through which “topical relationships, research trends, and complementary capabilities can be discovered” [23] (p. 351), allowing for the identification of the relevant sources and mapping of key concepts [24]. Such reviews are also sometimes termed “scoping reviews” [25] or “integrative reviews” [26], “with the aim to assess, critique, and synthesize the literature on a research topic in a way that enables new theoretical frameworks and perspectives to emerge” [26] (p. 335).

This review thus aimed to map recent developments in cross-modal fusion methods for audio–visual emotion recognition. Between May and September 2025, searches were conducted across IEEE Xplore, ACM Digital Library, Scopus, Web of Science, SpringerLink, and ScienceDirect, covering the period 2014–2025. Keywords included “cross-modal fusion”, “multimodal fusion”, “audio–visual deep learning”, “attention fusion”, “low-rank fusion”, “tensor fusion”, “tensor decomposition”, and “bilinear pooling”. This review focused on peer-reviewed studies aligned with audio, visual, multimodal, and cross-modal emotion recognition, particularly those employing deep learning, attention mechanisms, and low-rank tensor approximation techniques. Non-peer-reviewed works, unimodal studies, non-English papers, and studies without machine-learning-based fusion were excluded. A total of 178 records were identified, 28 duplicates were removed, 150 articles were screened, and 92 full-text papers were examined. Of these, 68 studies met the inclusion criteria, and these provided references to other studies, which were then assessed, which, in turn, provided further relevant sources. Studies of electrophysiological data from the brain (EEG) and heart (ECG) to detect emotional state were not considered. Such studies require specialized sensors and controlled environments, whilst this research focuses on dialog-based audio–video modalities that analyze external behavioral cues, such as facial expressions and vocal tone, for naturalistic emotion recognition. These represent distinct research domains with different objectives and data characteristics, and so EEG/ECG-based studies were intentionally excluded from the scope of the literature review.

An assessment of the literature supported the development of the PCF. A PCF not only gathers concepts but also integrates them into one single structure. The goal is to find factors, attributes, variables, behavior, processes, etc. that provide an initial analytical frame for subsequent primary research. According to Levering [27], a PCF is a good starting point to explain and examine a wider subject area. Similarly, Jabareen [28] suggested that a conceptual framework can provide an analytical network for investigating a particular phenomenon. The selection of concepts is based on the number of occurrences in a text, their meaning, and relevance to the research objectives or questions.

2.2. Experimental Design and Validation

Phase 2 focused on experimental development and evaluation of a novel cross-modal fusion using Tucker decomposition. Such experiments can be seen as "a method of gathering information and data on a subject through observation in controlled settings" [29]. Symbols and notations included in the equations and algorithms are shown in the abbreviation section at the end of this article. In emotion recognition, cross-modal fusion is motivated by the fact that emotional cues are distributed across multiple modalities. Relying on a single modality can lead to ambiguity, whereas combining audio and visual information enables the model to exploit complementary emotional signals for improved accuracy and robustness [30].

However, direct outer-product fusion of multimodal features produces a high-dimensional tensor with significant redundancy. To address this, Tucker decomposition [31] was employed, this being a powerful multilinear factorization technique that compresses the cross-modal tensor into a low-rank core tensor and learnable projection matrices. This allows the model to capture the most salient inter-modal correlations while maintaining computational efficiency. The Tucker framework provides flexibility in rank selection per mode and supports end-to-end learning, making it particularly suitable for fine-grained audio–visual emotion classification [30,32,33].

Cross-modal feature interaction and the associated feature extraction methods were used, combining information from two heterogeneous sources: audio and video frames used as modalities. Audio features capture vocal tone, referring to pitch, timbre, and modulation patterns that convey emotional cues such as anger, sadness, or happiness, while intensity reflects the loudness or energy level of speech, often associated with emotional arousal. In contrast, video frames convey facial expressions and gestures. Since audio and visual features represent distinct modalities, combining their features enables the model to capture richer and more complete information than using either alone in emotion recognition tasks. Appendix A provides details on the model architecture and hyperparamete configuration.

Figure 2 depicts the experimental design. Regarding audio feature extraction (left side in Figure 2), the 2D convolutional neural network processes the audio modality, specifically designed to extract high-level features while preserving their temporal sequence. This network operates on Mel-spectrogram inputs to learn emotion-relevant time–frequency patterns. The 2D convolutional layers apply small learnable filters that slide across the spectrogram to detect localized changes in pitch, energy, and spectral shape associated with emotional expression. Layers with different kernel sizes and strides capture patterns at multiple temporal and spectral scales, while batch normalization ensures training stability and accelerates convergence. The ReLU activation introduces non-linearity, allowing the model to learn complex mappings between acoustic cues and emotional states. Unlike models that collapse an entire clip into a single representation, this CNN preserves the sequential structure of the audio, producing a sequence of high-level features that reflect the evolution of emotion over time. The network comprises a series of 2D convolutional layers with varying kernel sizes and strides, followed by batch normalization and a non-linear activation function (ReLU). The detailed architecture is provided in Appendix A (Table A1). This architecture ensures that the output for each audio segment is not a single aggregated vector, but rather a high-dimensional audio feature sequence, as expressed in Equation (1):

$$F_{audio\_seq} \in \mathbb{R}^{T_{audio} \times D_{audio\_per\_step}} \tag{1}$$

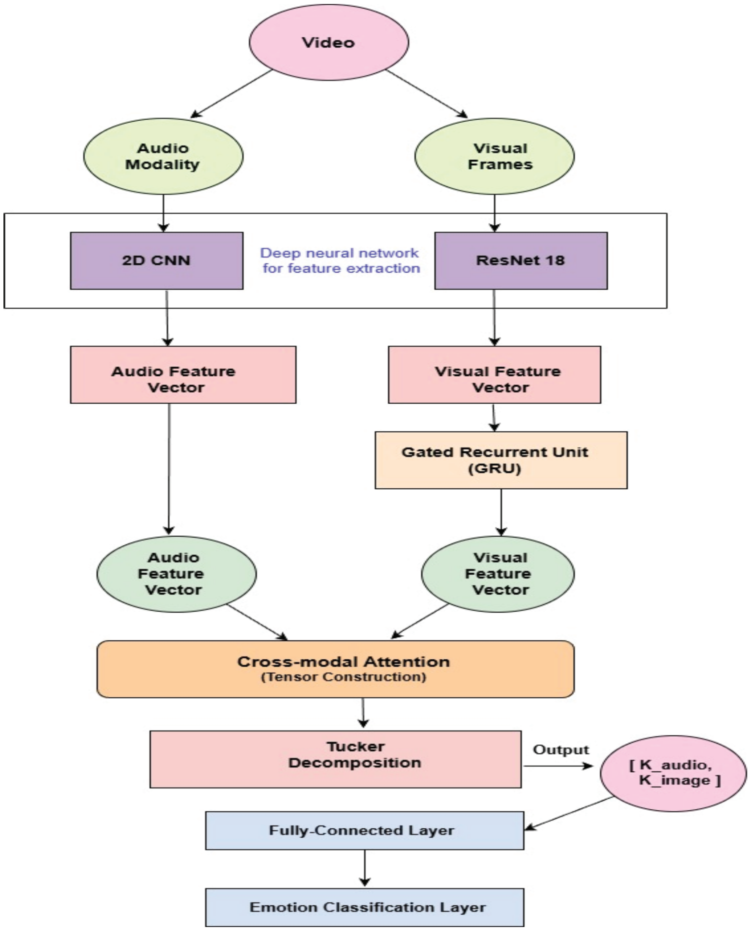

Figure 2. Experimental design and approach.

Here, T_{audio} represents the number of temporal steps (or frames) remaining after convolutions and pooling, and $D_{audio\_per\_step}$ represents the feature dimension at each of these time steps. The 2D convolutional neural network processes the audio modality, extracting high-level time–frequency features while preserving their temporal order. Raw audio segments are input as tensors of shape (batch size, sequence length), and the convolutional layers operate along both time and frequency axes. Rather than collapsing the entire segment into a single feature vector, the network outputs a sequence of frame-level feature embeddings. Preserving this temporal structure allows the model to represent variations in pitch, intensity, and rhythm are key aspects of emotional prosody that unfold over time. Such sequential representations are crucial for capturing dynamic emotional transitions and for synchronizing temporal cues with the corresponding visual frames in the subsequent fusion stage.

In terms of visual feature extraction (right side in Figure 2), individual video frames are processed as tensors of shape (batch size, height, width, channels). Batch size denotes the number of samples processed simultaneously during training or inference. Height and width correspond to the spatial dimensions of each feature map. Channels indicate the number of feature maps or filters, capturing different aspects of the input (e.g., color channels in RGB images or feature channels in convolutional feature maps). A pre-trained ResNet-18 model [34], initialized with ImageNet weights [35], is the backbone for extracting robust spatial features from each frame. The pre-training on the large-scale ImageNet dataset enables the network to learn generic low- and mid-level visual features (e.g., edges, textures, shapes), which are then fine-tuned on the emotion recognition dataset to adapt to domain-specific facial expression patterns.

The final classification layer of the pre-trained ResNet-18 model was removed because the objective is to extract discriminative visual features, rather than performing ImageNet classification. After removing the last fully connected (FC) layer, a 512-dimensional embedding from the global average pooling layer is obtained, which represents each input frame in a compact and informative manner. This feature vector encodes high-level spatial and semantic information (e.g., eye movement, mouth shape, and facial muscle tension), which is subsequently aligned and fused with the corresponding audio features for emotion classification. This approach enables transfer learning, where the lower layers capture general visual patterns while the final embedding reflects task-specific emotional cues.

The final classification layer of ResNet-18 is removed, allowing the network to output a high-dimensional visual feature vector for each frame. To prepare for temporal fusion, these frame-level features are then explicitly stacked along a temporal dimension to form a visual feature sequence for the entire video utterance, as shown in Equation (2):

$$F_{image\_seq} \in \mathbb{R}^{T_{image} \times D_{image\_per\_frame}}. \tag{2}$$

Here, T_{image} represents the number of frames in the utterance, and $D_{image\_per\_frame}$. It is the feature dimension extracted by ResNet-18 for each frame.

The sequence of frame-level visual features is subsequently processed by a Bidirectional Gated Recurrent Unit (Bi-GRU) to model temporal dependencies across the video frames. The Bi-GRU enables the network to capture how facial expressions evolve over time—for example, how a neutral expression gradually transitions into a smile or frown. By incorporating information from both past and future frames, the Bi-GRU learns contextual patterns such as intensity buildup, duration, and smooth emotional transitions, which cannot be inferred from isolated frames alone. This temporal modeling complements the spatial representations extracted by ResNet-18, thereby strengthening the system's ability to recognize emotions in continuous video sequences. Instead of using only the final hidden state, the sequence of hidden states produced by the Bi-GRU is retained to preserve frame-wise temporal context for downstream fusion. The resulting temporally refined visual feature sequence is denoted in Equation (3):

$$F'_{image\_seq} \in \mathbb{R}^{T_{image} \times D_{image\_per\_frame}} \tag{3}$$

This output represents the temporally refined visual feature sequence, which is subsequently used for cross-modal alignment and tensor construction. The cross-modal alignment step synchronizes the temporally refined visual features with their corresponding audio features at each common time step, ensuring consistent pairing across modalities. The aligned features are then used to construct a cross-modal tensor that captures fine-grained interactions between the two modalities for Tucker decomposition–based fusion (see Appendix B, Algorithm A2 for detailed steps and formulations).

Independent deep neural networks, namely a 2D CNN for audio and Resnet-18 for video, extract high-level, discriminative features from each modality. The deeper convolutional layers of ResNet-18 generate abstract, high-level representations that capture semantic and expression-specific cues, such as facial muscle activation patterns, eye and mouth region dynamics, and texture variations associated with different emotions. These features are termed "discriminative" because they enable the model to distinguish between visually similar emotional states (e.g., happiness vs. surprise) by emphasizing subtle yet meaningful spatial patterns. Moreover, these high-level visual embeddings facilitate cross-modal alignment with audio features during the fusion stage, as they encode semantically rich information that complements vocal attributes such as tone, intensity, and rhythm. This semantic alignment between modalities enhances the model's ability to form consistent and context-aware multimodal representations, leading to improved emotion recognition accuracy. In the proposed work, the extracted feature vectors from the audio $(F._{audio})$ and visual $(F._{image})$ modalities are fused using an outer product, which overcomes the limitations of linear fusion by modeling all pairwise correlations between audio and visual features. This results in a richer and more discriminative joint representation for emotion classification.

This proposed work enhances the fusion process by utilizing a cross-modal attention mechanism, which enables the model to dynamically learn which parts of one modality are most relevant when considering the other. The motivation for using this mechanism is to allow for context-aware feature refinement, where the model selectively emphasizes emotionally salient cues and suppresses noisy or less informative ones. This process helps ensure better temporal and semantic alignment between modalities before constructing the joint tensor for Tucker decomposition. It refines the modality-specific feature vectors, producing contextually richer representations: $(F'._{audio})$ and $(F'._{image})$. The model selectively emphasizes emotionally salient cues and suppresses noisy or less informative ones. The proposed model incorporates a cross-modal attention mechanism to facilitate interaction between the audio and visual modalities. This mechanism is designed to allow the model to learn the relevance of features across modalities, such that information from one stream can modulate the representation of the other. Through this process, the audio and visual feature vectors $(F'._{audio})$ and $(F'._{image})$ are contextually adjusted based on inter-modal correlations prior to fusion.

The attention operation serves to focus computational resources on the most informative regions in each modality, which theoretically contributes to more coherent multimodal feature representations before the subsequent Tucker decomposition. Additionally, the mechanism is structured to mitigate the influence of less informative or noisy features by adaptively weighting their contributions, thereby improving the stability of the fusion process.

This approach helps create a more contextually rich interaction before the subsequent Tucker decomposition, which is employed to efficiently model high-order correlations between the refined audio and visual features. The Tucker framework decomposes the cross-modal tensor into a low-rank core tensor and projection matrices, thereby capturing fine-grained interactions while mitigating the high dimensionality produced by the outer product fusion. This step provides a compact and expressive joint representation for emotion classification.

Another key advantage of the cross-modal attention mechanism is its ability to handle situations where one modality (e.g., audio or visual) becomes noisy or less informative. For example, in spontaneous or acted emotion datasets, visual cues may be compromised by poor lighting conditions, head movements, or partial facial occlusion, while audio signals may contain background noise, overlapping speech, or inconsistent vocal intensity.

In such cases, the attention mechanism adaptively down-weights the contribution of the unreliable modality, relying more heavily on the complementary one to preserve emotional consistency and enhance robustness. During training, attention coefficients are computed using a similarity function between the query and key representations, followed by a Softmax normalization that ensures the weights sum to one. Features associated with noisy or less informative inputs naturally receive lower attention scores, thereby reducing their influence in the fused representation. For example, when background noise affects an audio segment or a frame suffers from facial occlusion, the corresponding attention weights for that modality decrease, enabling the model to emphasize the cleaner, more informative modality. This dynamic weighting enhances both the robustness and interpretability of the emotion recognition process.

Moreover, the learned attention weights can provide interpretive insight into which features or modality interactions contribute most to the model's emotion recognition decision. Attention weights do not offer full interpretability, but can provide indicative insights into the regions or modalities the model focuses on during decision-making. Such qualitative observations can help identify which audio–visual interactions are emphasized during emotion recognition, offering a limited but useful perspective on the model's internal behavior.

In the proposed cross-modal attention mechanism, the sequential features from one modality act as the Query (Q), while those from the other modality serve as Key (K) and Value (V). Rather than merely exchanging information, this mechanism enables the model to compute a relevance score between every element of the two modalities using a similarity function (e.g., the dot product), which is then normalized through a Softmax function to produce attention weights. These weights determine how strongly each element in one modality should attend to the other, allowing the model to emphasize emotionally consistent cues and suppress irrelevant or noisy signals. The process is performed in both directions, with visual attention to audio and audio attention to visuals, to generate context-enriched feature sequences that are temporally aligned and semantically refined before the Tucker decomposition stage. A detailed description of this mechanism, including attention head configuration and embedding dimensions, is provided in Appendix A.3.

The audio feature sequence $\left(F_{audio\_seq}\right)$ acts as the Query (Q) and attends to the visual feature sequence $\left(F_{image\_seq}\right)$ (Key/Values), producing the attention-weighted visual feature representation (F'_{image_seq}). Here, "querying" means that each audio time-step feature evaluates the relevance of all visual frame features to determine which visual cues correspond most closely to that segment of audio (e.g., matching voice pitch or tone with facial expression). The attention mechanism computes these similarity scores using a dot-product between query and key representations, normalized through Softmax, producing attention weights that emphasize the most relevant visual features for each audio step. This allows the model to highlight visual cues and determine which ones correspond most closely to a specific segment of audio (e.g., matching voice pitch or tone with facial expression). The model computes attention scores that represent the similarity between each visual frame and all audio time-step features. These scores quantify the strength of the relationship between each visual feature and each audio feature, based on learned semantic and temporal correlations.

Conversely, the visual feature sequence $F_{\text{image\_seq}}$ serves as the Query and attends to the audio feature sequence $F_{\text{audio\_seq}}$, generating the attention-weighted audio representation $F'_{\text{audio\_seq}}$. The model computes attention scores that represent the similarity between each visual frame and all audio time-step features.

This process allows every visual time-step to selectively emphasize the most relevant portions of the audio sequence, for instance, associating lip movements or facial tension with corresponding pitch, intensity, or rhythm variations. In this manner, the audio representation is refined in a context-aware and temporally aligned manner before Tucker decomposition.

The attention mechanism in the model is designed to re-weight features across time and modality without altering their dimensional structure. Maintaining the same temporal (T_{common}) and feature (D_{audio}, D_{image}) dimensions ensures that:

- Temporal alignment between audio and visual modalities is preserved, allowing for one-to-one correspondence across time steps.
- Tensor construction for Tucker decomposition remains consistent, since outer product fusion requires matching time-step alignment, and
- Feature interpretability and spatial integrity are retained, avoiding information loss from dimensional projection.

These attention-weighted feature sequences $\left(F'_{audio\_seq} \text{ and } F'_{image\_seq}\right)$ retain the original sequential and feature dimensions because the attention mechanism modifies the importance weighting of features rather than their dimensionality. Preserving these dimensions ensures that both modalities remain temporally aligned (T_{common}) and structurally compatible for subsequent tensor construction in the Tucker decomposition stage. This design allows the model to encode inter-modal dependencies while maintaining the temporal order and feature-space integrity of each modality, enabling accurate outer-product fusion without loss of correspondence. Specifically, the refined audio and visual sequences are defined in Equations (4) and (5):

$$\left(F'_{audio\_seq} \in \mathbb{R}^{T_{common} \times D_{audio\_per\_timestep}}\right) \tag{4}$$

$$\left(F'_{image\_seq} \in \mathbb{R}^{T_{common} \times D_{image\_per\_frame}}\right) \tag{5}$$

However, the attention mechanism dynamically learns and updates its weighting values based on the similarity between feature representations from the two modalities. At each time step, attention scores are computed as the dot-product similarity between the query and key projections and are normalized using Softmax to form a probability distribution over all possible interactions. These weights are optimized during training, allowing the model to gradually emphasize strongly correlated audio–visual pairs (e.g., raised pitch aligned with smile intensity) and de-emphasize weak or irrelevant ones. Consequently, the attention mechanism highlights inter-modal correlations and captures contextual relevance over time, producing refined feature sequences that are then fused to construct the higher-order tensor for Tucker decomposition.

Tucker decomposition is applied to this cross-modal tensor to reduce redundancy and computational complexity arising from the high-dimensional outer-product representation. The model isolates the most informative latent components that capture essential cross-modal correlations by decomposing the tensor into a low-dimensional core tensor and factor matrices while discarding irrelevant or redundant variations. This compression not only yields a compact and discriminative joint representation but also facilitates end-to-end optimization as a learnable layer within the network, enhancing both efficiency and generalization for the emotion recognition task (see Appendix A for implementation details).

The core tensor encapsulates the most salient and statistically significant audio–visual interactions by jointly modeling temporal, spectral, and spatial dependencies. This compact latent representation has been shown to be highly discriminative and informative for affective tasks because it retains the strongest cross-modal correlations while filtering redundant noise [36]. Empirically, the improved performance of the model over baseline fusion methods further supports that the Tucker-derived core serves as a representative joint feature for emotion recognition. All components of the proposed architecture are jointly trained in an end-to-end manner using backpropagation, allowing gradients to flow through the feature extraction, cross-modal attention, and Tucker fusion layers. The network parameters are updated iteratively using the categorical cross-entropy loss computed over emotion labels, ensuring that the model learns discriminative representations directly optimized for the emotion classification task.

The flexibility of Tucker decomposition is leveraged to enable a richer tensor representation beyond simple vector fusion (Figure 3). In previous studies on multimodal fusion, such as the Tensor Fusion Network [30], MUTAN [32], and BLOCK Fusion [37], the batch dimension of the input tensor $\left(A^{(Batch)} \text{ or } A^{(N)}\right)$ is typically preserved but not explicitly decomposed. These approaches perform per-sample cross-modal fusion, where each tensor represents the interactions among modalities for an individual sample while maintaining batch processing for parallel training efficiency. These approaches perform per-sample cross-modal fusion, where each tensor represents interactions among modalities for an individual sample, while batch processing is maintained for parallel training efficiency.

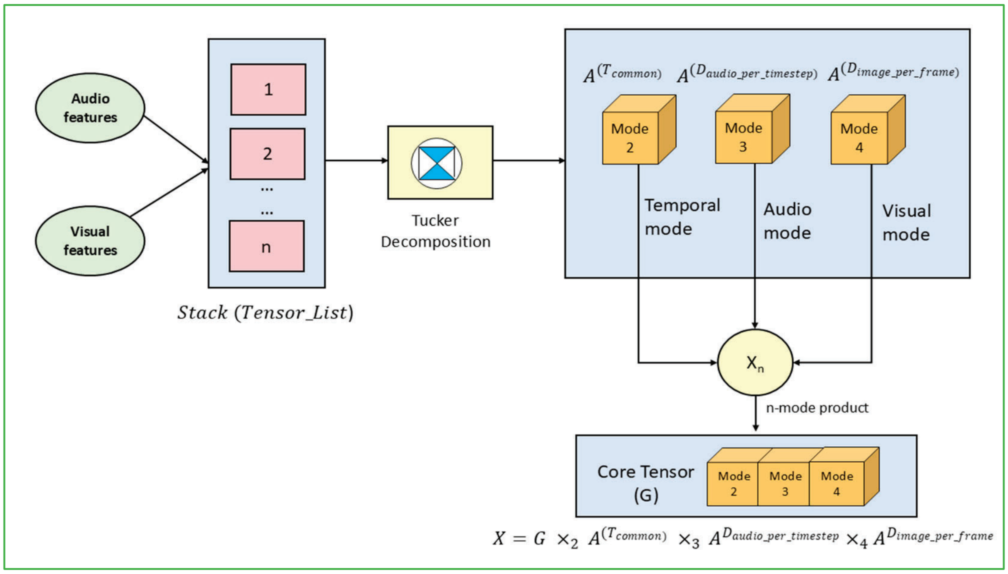

Figure 3. Cross-modal fusion using the Tucker decomposition.

These approaches perform per-sample cross-modal fusion, where each tensor represents interactions among modalities for an individual sample, while batch processing is maintained for parallel training efficiency. The primary goal is to find a shared, lower-dimensional representation for the interaction between modalities within each sample. However, the proposed cross-modal tensor construction in this study is distinct in that it integrates attention-weighted audio and visual feature sequences through a time-aligned outer-product operation, forming a high-order tensor that explicitly retains both temporal and feature modes of each modality. This design captures fine-grained spatiotemporal dependencies beyond simple vector concatenation. The initial input tensor, X, is defined with more modes directly relevant to the temporal and feature dimensions of audio and visual data, enabling a more granular and powerful fusion. The detailed formulation and step-by-step construction process for the tensor are provided in Appendix B.

Extensive experiments are conducted on three widely used datasets- IEMOCAP [38], RAVDESS [39], and CREMA-D [40] to validate the effectiveness of the fusion framework. Profiles of these datasets are provided in Appendix C. All experiments were conducted on a high-performance system detailed in Appendix D. The effectiveness of the proposed cross-modal fusion framework was evaluated using the standard evaluation metrics: precision, recall, F1-score, and accuracy, as formally defined in Equations (6)–(9).

Precision: Precision measures the proportion of correctly predicted positive instances among all instances predicted as positive. High precision indicates that the model makes few false positive errors when predicting the positive class.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False positive}} \tag{6}$$

Recall (Sensitivity): Recall measures the proportion of correctly predicted positive instances among all actual positive instances. High recall indicates that the model successfully detects the true positive instances.

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \tag{7}$$

F1 Score: F1 Score is a reliable metric for evaluating classification performance, particularly when the dataset has imbalanced class distributions. F1 score provides a more balanced measure of a model's accuracy by combining precision and recall.

$$\text{F1 Score} = \frac{2 \, * \, \text{Precision} \, * \, \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

Accuracy: Accuracy provides a straightforward indicator of the model's consistency throughout all classes.

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \tag{9}$$

3. Results

This section presents the main results of this study, directly addressing the two RQs in Sections 3.1 and 3.2. Other emergent issues are discussed and analysed in Section 4.

3.1. RQ1. To What Extent Do Affective Computing and Recent Advancements in Deep Learning Improve Cross-Modal Fusion for Audio and Image-Based Emotion Recognition?

Affective computing—which combines elements of computer science, psychology, and cognitive science—enables the development of computing devices that can analyse voice, text and facial expressions, thereby comprehending emotional states and customising responses accordingly. Previous studies in the field primarily focused on unimodal approaches [41–44], leveraging either audio or image data independently for emotion recognition. These traditional methods, while effective to an extent, often lacked robustness due to their reliance on a single modality. Recent research interest has progressively shifted toward cross-modal emotion recognition [45–48], where two or more modalities are integrated to enhance recognition accuracy and contextual understanding. This evolution has been underpinned by advances from conventional machine learning to more sophisticated deep learning architectures in recent years. Based on an analysis of the extant literature, recent progress on audio and image-based fusion mechanisms is assessed, focusing on cross-modal alignment designed to capture complementary spatiotemporal information from both modalities.

The extraction of audio signal features has seen various advancements since the mid-20th century. The earliest and most fundamental features were extracted from the time domain, with notable progress continuing until the late 1950s [49]. Time-domain features, such as the Zero Crossing Rate [50], Root Mean Square (RMS) Energy, and Waveform

Shape [51], provided foundational tools for analyzing audio signals. Researchers then began exploring the frequency domain, introducing features like spectral centroid, bandwidth, contrast, flatness, and roll off [52] from the spectrograms [53], which offered insights into the spectral characteristics of signals.

To address the limitation of both time and frequency domain approaches, a joint time-frequency [54], and domain features such as Mel-Frequency Cepstral Coefficients (MFCCs) [55], Chroma Features, Mel Spectrogram, and Constant-Q Transform (CQT) [56], were developed. These features enabled a more prominent representation of audio signals by capturing temporal and spatial dynamics [57]. Additionally, advanced features like Pitch and Harmonics [8], Formants, and Temporal Modulation [58] explored and expanded statistical capabilities, supporting more complex applications in audio signal processing.

Despite these advancements, some limitations persist. For example, while time-domain features such as short-term energy, zero-crossing rate, or waveform amplitude are computationally efficient, they often fail to capture frequency-specific cues like pitch contour, harmonic structure, and formant shifts elements which strongly correlate with emotional tone and intensity [59]. Conversely, frequency-domain representations such as spectrograms, MFCCs, or chroma features effectively capture these acoustic dynamics but are computationally more demanding, making them less suitable for real-time or resource-constrained applications. These trade-offs continue to drive innovation in audio feature extraction for emotion recognition tasks.

Feature extraction from video data involves processing spatial and temporal information to capture meaningful patterns. Spatial features [57] are typically extracted frame-by-frame using convolutional neural networks, identifying objects, textures, and expansions [60]. Temporal features [58], crucial for understanding motion and dynamics, are captured by optical flow [61] or any deep neural network [62]. The process poses several challenges, including the high computational cost of processing large volumes of frames and maintaining temporal coherence.

Facial feature extraction is a key challenge and plays a key role in emotion recognition. It leverages some techniques, such as the Facial Action Coding System (FACS) [63], handcrafted features [64], and facial expression features [65]. Handcrafted features have garnered significant attention due to their effectiveness in representing facial features. Some traditional techniques include Gabor filters [66], which are adept at capturing texture and orientation information; Scale-Invariant Feature Transform (SIFT) [67], which excels in identifying distinctive key points; Local Binary Patterns (LBP) [68], renowned for their texture representation capability.

The feature extraction step is critical in reducing the dimensionality of input data, thereby enhancing model accuracy and mitigating the risks of overfitting. By efficiently extracting meaningful patterns, these techniques lay the foundation for the emotion recognition task. However, handcrafted features often struggle to generalize across diverse datasets, as their performance may depend heavily on several environmental factors, and overall, they struggle to capture subtle facial expressions.

Thus, fusion mechanisms for cross-modal emotion recognition are required to integrate features from multiple heterogeneous data sources or modalities (Figure 4). This approach is particularly advantageous in scenarios where information from various domains must be combined to understand the underlying patterns or behaviors comprehensively. The fusion process typically involves several strategies: early fusion [18], late fusion [20], hybrid fusion [21], model-level fusion [69], hierarchical fusion [70], rule-based decision-level fusion [71], and estimation-based fusion [72]. These fusions are mainly subdivisions of Intra-modal fusion and Cross-modal fusion [45]. Early fusion combines features from different modalities at the initial stages, creating a unified feature set before applying machine learning algorithms. Ortega et al. [73] proposed an approach using early fusion; the study emphasizes the importance of meticulous hyperparameter optimization to address challenges such as premature overfitting. Conversely, late fusion processes each modality separately and then merges their outputs at the decision level [74]. Hybrid fusion combines early and late fusion elements, integrating features at multiple stages to capitalize on their advantages [75]. Model-level fusion involves combining different models trained on separate modalities by averaging their outputs or using more sophisticated techniques such as stacking or boosting [69].

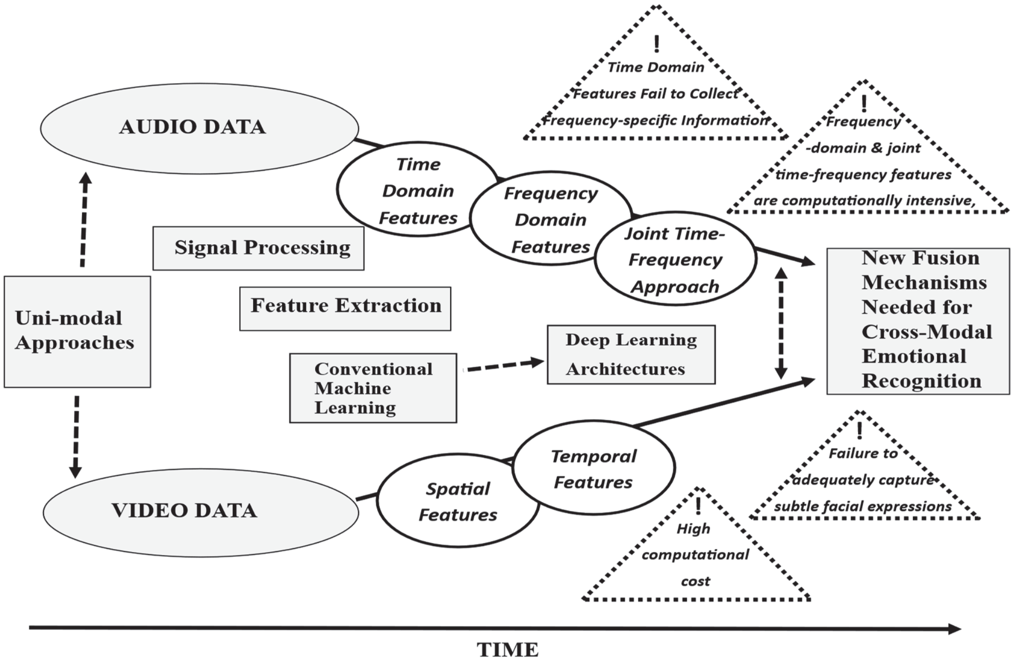
AUDIO DATA
Time Domain Features
Frequency Domain Features
Joint Time-Frequency Approach
!
Time Domain Features Fail to Collect Frequency-specific Information
!
Frequency -domain & joint time-frequency features are computationally intensive,
Signal Processing
Feature Extraction
Uni-modal Approaches
Conventional Machine Learning
Deep Learning Architectures
New Fusion Mechanisms Needed for Cross-Modal Emotional Recognition
VIDEO DATA
Spatial Features
Temporal Features
!
Failure to adequately capture subtle facial expressions
!
High computational cost
TIME

Figure 4. Cross-modal emotion recognition: conceptual framework. The figure depicts the evolution of fusion mechanisms for audio and video data from uni-modal to cross-modal approaches, indicating technology advancements (shaded grey), new features and limitations (!).

Hierarchical fusion employs a multi-layered approach, where data is fused at various levels of abstraction, allowing for a more nuanced and detailed information integration [76].

The cross-modal fusion mechanism [75,77,78] effectively integrates and interacts with contextual relevance among features from different modalities. These existing works provided the foundation for the design of an interpretability method for audio and video-based emotion recognition to justify the significance of every input visual characteristic and every vocal segment. By leveraging the complementary strengths of different data types, fusion mechanisms enhance robustness and reliability. This approach is extensively applied for emotion recognition, but also in other fields including medical diagnostics and autonomous driving, where integrating diverse data sources results in more accurate and insightful outcomes. Recent research in cross-modal emotion recognition has focused on using cross-attention feature-level techniques to improve the merging of audio and image features. Praveen et al. [78] developed a joint cross-attention model that effectively integrates audio and visual cues for dimensional emotion detection, allowing for a more comprehensive understanding of complicated emotional states. Similarly, Mocanu et al. [45] demonstrated how cross-modal audio-video fusion, attention processes, and deep metric learning improved identification accuracy by aligning modality-specific representations. The study by Zhou et al. [79] described a cross-attention and hybrid feature weighting neural network, emphasizing the importance of balanced feature contributions in emotion recognition from large-scale video clips.

Furthermore, Lee et al. [80] investigated speech emotion identification using cross-modal fusion methods that align audio and textual information, demonstrating the synergy of linguistic and acoustic modalities. These developments demonstrate cross-modal effectiveness in capturing detailed interactions across modalities, overcoming the limitations of cross-modal data fusion, and improving emotion recognition ability. Despite these advances, limitations persist. Transformer-based models [81] frequently require significant processing resources, which might be prohibitive for real-time applications. At the same time, relying on large-scale and balanced datasets for training may result in lower performance in domains with limited or imbalanced data. These problems underscore the importance of the search for new efficient and scalable techniques for cross-modal emotion identification.

3.2. RQ2. Can a New Model Be Developed and Validated to Improve Correlations Between Heterogeneous Modalities (Such as Audio and Visual Data) to Detect and Interpret Human Emotional States?

This section presents the experimental results of the cross-modal fusion mechanism for emotion recognition using the Tucker decomposition framework, demonstrating its effectiveness across three datasets: RAVDESS, CREMA-D, and IEMOCAP (see Appendix C). The performance is evaluated using standard classification metrics, including precision, recall, and F1-score, with macro-averaged and per-emotion. Table 1 illustrates the per-emotion performance of the model on the RAVDESS dataset, which includes 8 distinct emotional classes.

Table 1. Model performance on the RAVDESS dataset.

Emotions	Angry	Calm	Disgust	Fearful	Happy	Neutral	Sad	Surprise
Precision	92.12	92.8	91.47	93.2	93.85	91.47	91.17	93.6
Recall	92.8	93.91	91.64	91.06	93.12	93.5	91.55	91.55
F1-Score	92.46	93.35	91.55	92.12	93.49	92.47	91.36	92.56

The model achieves consistently high performance across all emotions on RAVDESS, reflecting its robust ability to distinguish between a wide range of expressive states: in particular, “Happy”, “Calm”, and “Surprise” are recognized with effective accuracy. Even the lowest F1-score for “Sad” remains remarkably high, indicating a well-balanced classification capability without significant bias towards specific emotions. This balanced performance illustrates the effectiveness of the approach to Tucker decomposition in leveraging complementary cues from audio and video modalities.

Table 2 illustrates the model’s strong performance on the CREAM-D dataset. “Disgust” achieves the highest F1-score, closely followed by “Fear” and “Happy”, while “Neutral” and “Anger” show slightly lower, yet still very competitive, predictions. The overall high F1-score across all emotions confirms the model’s ability to effectively generalize and classify emotions in this dataset.

Table 2. Model performance on the CREMA-D dataset.

Emotions	Anger	Disgust	Fear	Happy	Neutral	Sad
Precision	86.5	88.8	87.93	87.39	85.62	85.62
Recall	85.23	88.46	87.4	87.83	85.08	88.88
F1-Score	85.86	88.63	87.67	87.61	85.35	87.22

Table 3 illustrates the model's robust performance on the more complex IEMOCAP dataset. F1-scores for "Anger" and "Frustration" are among the best predicted emotions, highlighting the model's capability to handle highly distinct and subtle cue-based emotional states. While "Excited" shows the lowest F1-score, it still represents a strong performance given the inherent variability and often subtle nature of this emotion in spontaneous audio waves. The overall balanced F1-Scores across the diverse set of emotions in IEMOCAP further validate the model's effectiveness in capturing complex cues.

Table 3. Model performance on the IEMOCAP dataset.

Emotions	Anger	Happiness	Sadness	Neutral	Excited	Frustration	Surprise	Fear
Precision	83.87	84.22	82.85	83.14	81.38	85.87	84.75	82.62
Recall	85.08	82.42	84.45	84.64	82.11	82.89	82.41	83.94
F1-Score	84.47	83.31	83.64	83.88	81.74	84.36	83.56	83.27

Table 4 provides a comprehensive result of the model's overall evaluation performance, presenting both weighted and unweighted accuracy metrics across the three datasets. Weighted accuracy accounts for class imbalance by averaging the accuracy of each class, while unweighted accuracy is simply the total number of correct predictions divided by the total number of samples. The result in Table 4 illustrates the model's robust performance. The consistently high weighted accuracy scores across all datasets confirm the model's ability to perform well even in the presence of potential class imbalances, highlighting its balanced classification capability for each emotion class.

Table 4. Model performance evaluation: weighted and unweighted accuracy metrics on RAVDESS, CREMA-D, and IEMOCAP datasets.

Dataset	Weighted Accuracy	Unweighted Accuracy
RAVDESS	92.46	88.74
CREMA-D	87.31	84.27
IEMOCAP	83.22	79.63

To validate the efficacy of the cross-modal fusion framework, a comprehensive analysis was conducted against various baseline fusion methods and several state-of-the-art cross-modal emotion recognition approaches. Table 5 presents a detailed performance comparison, specifically focusing on the RAVDESS dataset, a key benchmark for this domain.

Table 5. Performance comparison with other state-of-the-art cross-modal approaches on the RAVDESS dataset.

Source	Fusion	Accuracy	Remarks
[82]	Self-attention	75.76	The analysis relies solely on one dataset
[83]	Late fusion	86.70	The analysis relies solely on one dataset
[46]	Concatenation	66.90	Simple concatenation, high dimensionality
[45]	Cross-Attention	89.25	Poor sensitivity to micro-expressions
[84]	Cross-Attention	82.42	Computationally expensive and performed on a single dataset
Proposed Model	Cross-modal attention	92.46	Reduces dimensionality while employing three established benchmarks

Table 5 demonstrates superior performance by the proposed model on the RAVDESS dataset, achieving an accuracy of 92.46%. This significantly surpasses other contemporary methods, highlighting the effectiveness of the cross-modal attention fusion mechanism and Tucker decomposition-based multi-linear fusion. For instance, approaches utilizing self-attention [82] or simple concatenation [46] yielded substantially lower accuracies of 75.76% and 66.90%, respectively, often due to limitations such as reliance on single datasets or inherent high-dimensionality challenges. This comparative analysis indicates distinct advantages of the proposed model in the field of cross-modal emotion recognition.

Tables 6 and 7 illustrate that the model consistently demonstrates superior performance across both the CREMA-D and IEMOCAP datasets when compared to other state-of-the-art cross-modal approaches. On CREMA-D, the model achieved an accuracy of 87.31%, surpassing methods like those set out by Mocanu et al. [45] and Goncalves et al. [47], and significantly outperforming speaker-dependent models such as that of John and Kawanishi [48]. This highlights the effectiveness of the approach put forward here in capturing common temporal dynamics and leveraging the GRU layer for robust feature learning on this dataset. Similarly, on the highly challenging IEMOCAP dataset, the model secured an accuracy of 83.22%. This result is notably higher than that achieved by Moorthy and Moon [85]. The consistent outperformance on IEMOCAP underscores the critical role of our framework's intra-modal temporal refinement and sophisticated cross-modal attention fusion mechanism using Tucker decomposition in handling the complexities of spontaneous, cross-modal emotional expressions.

Table 6. Performance comparison with other state-of-the-art cross-modal approaches on the CREMA-D dataset.

Source	Fusion	Accuracy	Remarks
[48]	Multi-branch attention	72.45	Speaker-dependent model
[45]	Cross Attention	84.57	Poor sensitivity to micro-expressions
[47]	Conformer encoder	77.9	Versatile learning model
Proposed Model	Cross-modal attention	87.31	Common temporal dynamics along with the GRU layer

Table 7. Performance comparison with other state-of-the-art cross-modal approaches on the IEMOCAP dataset.

Source	Fusion	Accuracy	Remarks
[85]	Hybrid Multi-Attention Fusion	75.39	Parallel co-attention mechanism
Proposed Model	Cross-modal attention	83.22	Intra-modal temporal refinement

The confusion matrices are presented in Figures 5–7 for the RAVDESS, CREMA-D, and IEMOCAP datasets. These provide a granular insight into the model's per-emotion classification performance. These matrices visually complement the aggregate accuracy metrics by showing the number of correctly classified instances for each emotion (represented by strong diagonal elements) and identifying any specific emotions that are frequently confused with others (off-diagonal elements).

The lower accuracy observed on the IEMOCAP and CREMA-D datasets compared to RAVDESS primarily stems from the inherent variability and complexity of these datasets. RAVDESS consists of acted, high-quality, and well-balanced recordings captured under controlled conditions, resulting in clearer emotional cues and consistent expressions across subjects. In contrast, IEMOCAP and CREMA-D contain spontaneous, heterogeneous samples with variations in speaker identity, emotional intensity, accent, and recording conditions, which increase the difficulty of emotion classification. Figure 5 represents the confusion matrix for the RAVDESS dataset, illustrating class-wise performance of the proposed model on that specific dataset.

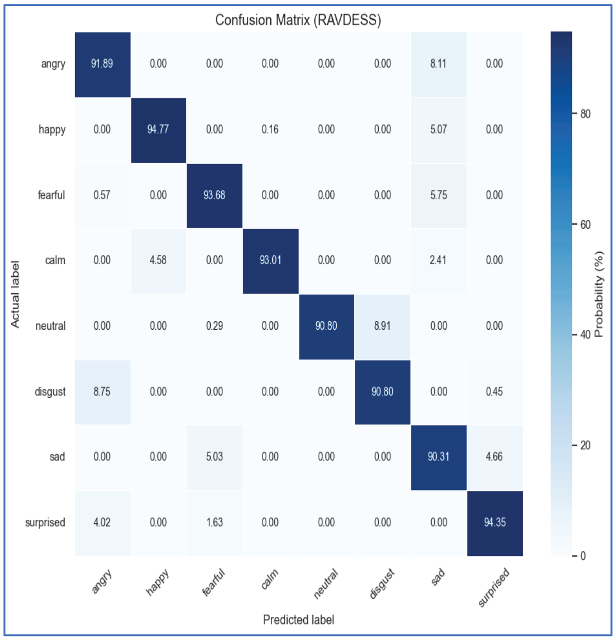

Figure 5. Confusion matrix of the RAVDESS dataset.

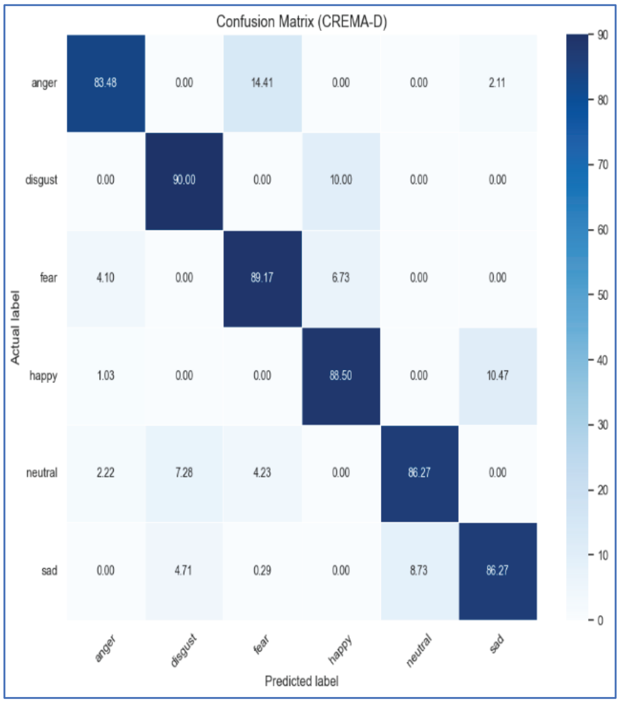

Figure 6. Confusion matrix of CREMA-D dataset.

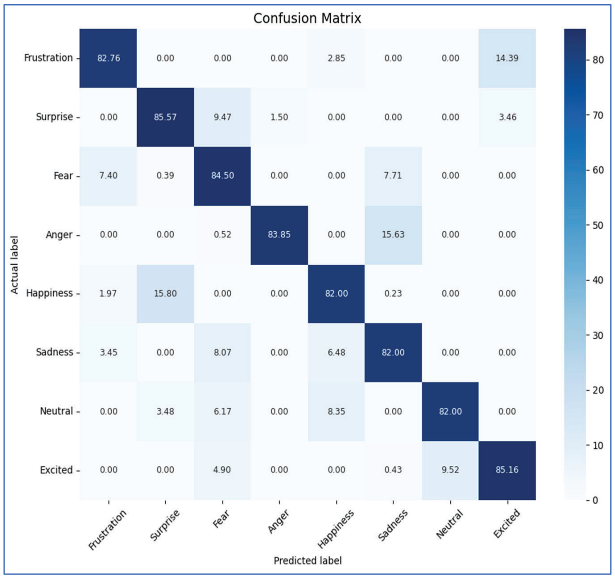

Figure 7. Confusion matrix of the IEMOCAP dataset.

4. Discussion

The above results raise three related issues worthy of further analysis and discussion. First, as regards the audio processing segment for emotion recognition in the model, not all parts of the spectrogram are equally important. Figure 8 provides a visualization of cross-attention weight matrices between audio and visual modalities. Each heatmap shows how visual frame features (queries) attend to audio frame features (keys/values). The horizontal axis represents audio time steps, the vertical axis visual frames, and the color intensity denotes normalized attention weights. The left panel depicts randomly distributed weights before convergence (no clear temporal structure); the middle panel shows focused attention along the diagonal, indicating temporal alignment between modalities; and the right panel illustrates selective peaks corresponding to key emotional moments. These visualizations qualitatively demonstrate how the attention mechanism tends to concentrate on temporally or contextually relevant segments, rather than uniformly attending across the entire sequence.

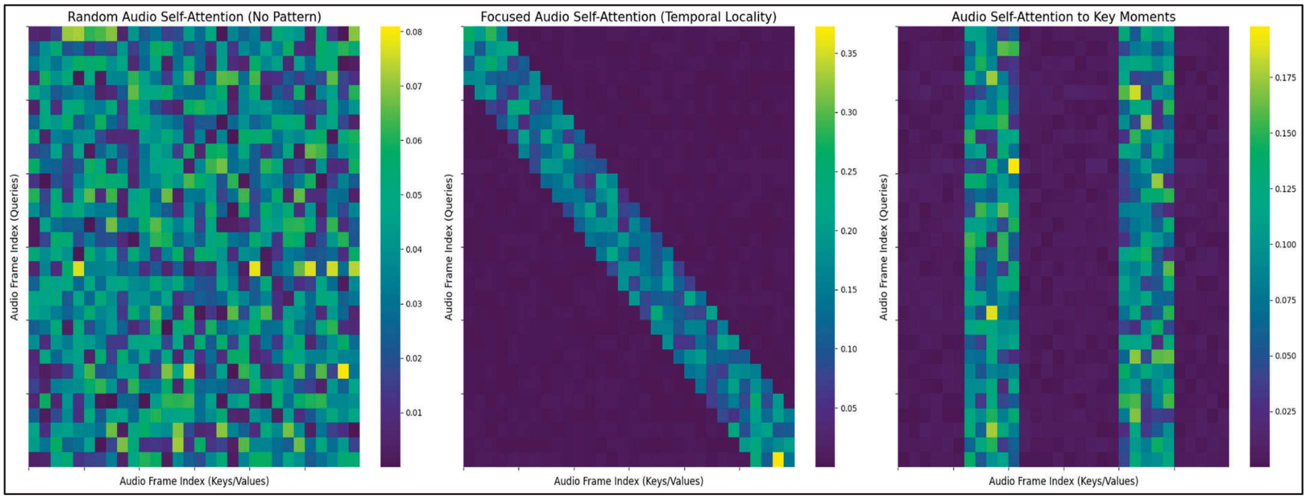

Figure 8. Visualization of attention weights within the audio modality.

Darker shades indicate higher attention scores, highlighting specific-time frequency regions that are more silent for the model's emotion recognition.

Figure 9 demonstrates the approach of the attention mechanism in video data. Unlike static images, videos contain a temporal dimension, requiring the model to identify both spatially and temporally salient regions. Relevance is determined through the attention weights, which are computed by measuring the dot-product similarity between the query and key feature representations for each frame. These scores are then normalized using a Softmax function to assign higher weights to frames or regions that exhibit stronger correlations with the emotional context. The resulting visualization therefore highlights the areas or segments that the model assigns higher attention weights to during emotion inference.

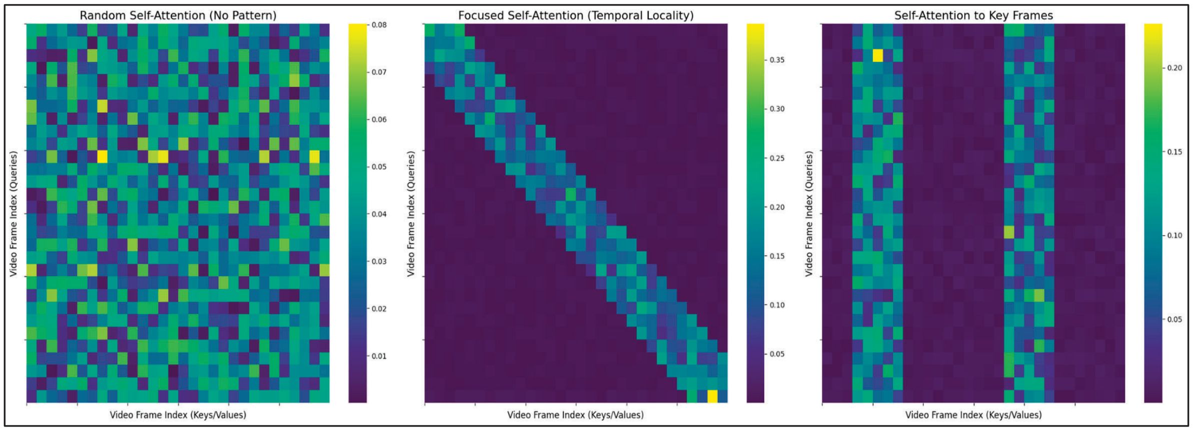

Figure 9. Visualization of attention weights within the video modality.

Brighter regions indicate higher attention, showing the model's focus on particular spatial areas and temporal frames within a video sequence.

Figure 10 illustrates the cross-modal attention mechanism that allows elements from one modality to attend to elements in another, revealing inter-modal dependencies. This visualization is important for understanding the cross-modal integrated information and resolving ambiguities by leveraging complementary cues from diverse data sources, ultimately leading to more robust and accurate predictions.

Secondly, and more specifically, the results provided examples of audio-visual cross-attention, where the flow of attention is from the visual modality to the audio modality. This means that features extracted from the visual stream are "querying" or attending to specific parts of the audio stream to enhance their understanding. For example, in a scenario where a person is speaking, the visual representations of their lips might attend to the corresponding speech sounds in the audio track. This type of attention helps the model to establish correspondence between what is seen and what is heard, allowing it to identify temporal alignments or causal relationships. It is particularly useful in refining visual interpretations.

In audio–video emotion recognition, the query–key–value (QKV) mechanism in the attention model enables each modality to selectively attend to the most relevant features in the other, effectively integrating complementary emotional cues across audio and video. Modality-specific encoders (CNNs for video and audio) extract high-level features, including facial expressions, speech tone, and rhythm. During cross-modal attention, the feature sequence from one modality (for example, audio) serves as the Query (Q), while the other modality (video) provides the Keys (K) and Values (V). The attention module computes relevance scores via dot-product similarity between Q and K, and uses these scores to weight the Values. The resulting weighted features highlight where emotional information in one modality aligns most strongly with the other. The mechanism operates bidirectionally: audio features attend to visual features, and visual features attend to audio. This bidirectional exchange enables the model to associate prosodic variations (pitch, tone, rhythm) with facial expressions and muscle dynamics, yielding richer multimodal representations for emotion understanding.

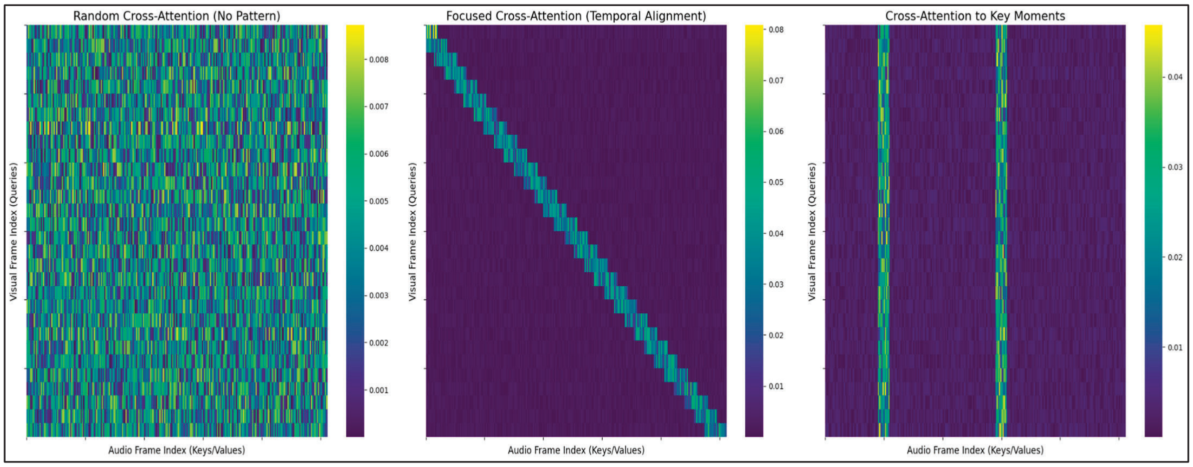

Figure 10. Visualization of cross-modal attention between audio and video modalities.

The cross-modal attention mechanism contributes three important functional advantages within the proposed framework. First, it facilitates temporal synchronization by aligning audio and visual cues that may express the same emotion slightly out of phase—for example, a shift in speech tone often occurs just before a corresponding facial expression, such as a smile or frown. Second, it performs feature selection, automatically assigning higher weights to emotionally salient patterns such as variations in voice pitch, vocal intensity, or facial muscle tension while suppressing background noise or frame-level distractions. Third, it enhances robustness by emphasizing cross-modal signals that consistently co-occur across audio and visual streams, allowing the model to remain effective even when one modality contains missing, degraded, or ambiguous information.

Together, these properties help the attention mechanism generate more coherent and context-aware multimodal representations prior to the Tucker fusion stage. Overall, cross-modal attention provides a data-driven way to connect vocal prosody and facial expressions, thereby producing semantically aligned, discriminative features before the Tucker-decomposition-based fusion stage [86,87].

The results also illustrate how, in the context of visual-audio cross-attention, the audio modality is attending to the visual modality, meaning that each audio feature vector computes attention weights over all visual frame features to identify which visual cues are most relevant to that audio segment. Here, the audio features "query" the visual information to gain a richer context. In the context of audio-to-visual cross-attention, the audio modality actively attends to relevant visual cues to gain contextual information. In this mechanism, each audio feature vector acts as a Query (Q) that computes a similarity score with all visual feature vectors (Keys, K) to identify which visual frames are most relevant to the current sound segment. The resulting attention weights are then used to form a weighted sum of the corresponding visual features (Values, V), producing an audio representation enriched with visual context. For example, when processing a low-pitched, slow vocal tone associated with sadness, the model may attend to visual cues such as drooping eyelids or a downturned mouth to reinforce its emotional interpretation. This cross-attention process enables the model to resolve ambiguities in vocal cues and recognize nuanced emotional states more accurately. In other words, the model learns to associate specific acoustic patterns, such as changes in pitch, tone, or intensity, with corresponding facial expressions or movements. This process produces an attention-weighted visual context that is fused back into the audio representation, allowing the model to interpret vocal emotions with the aid of visual cues. For instance, when a sad tone is detected, the audio features may focus more on visual indicators of sadness (e.g., drooping eyelids, downturned mouth), helping the model to disambiguate or refine emotional interpretation.

This attention mechanism allows the model to leverage visual cues to disambiguate or enhance the interpretation of emotional vocalizations. It is vital in identifying the emotion where visual context is crucial for understanding nuanced emotions, such as distinguishing genuine and feigned emotional expressions, or interpreting subtle emotional states, where audio cues alone might be ambiguous.

Thirdly, a deeper understanding of different extracted features can be displayed in a feature correlation heatmap (Figure 11). This visualization indicates the relationships and interdependencies between these features, whether they are acoustic (pitch, intensity, formants) or visual (facial landmarks, mouth openness). Each cell in the heatmap represents the Pearson correlation coefficient between two specific features. A high positive correlation, represented by warmer colours like red or orange, suggests that as the value of one feature increases, the other tends to increase as well. Conversely, a high negative correlation, represented by cooler colours like blue or dark blue, indicates that as one feature increases, the other tends to decrease. A value close to zero is often represented by white or light grey, indicating little to no linear relationship. Analyzing this heatmap helps the identification of redundant features, highly influential features, and potential dependencies that can inform feature selection, model architecture design, and improve the accuracy of emotion recognition. For instance, a strong correlation between vocal pitch and eyebrow movement might suggest a shared underlying emotional expression.

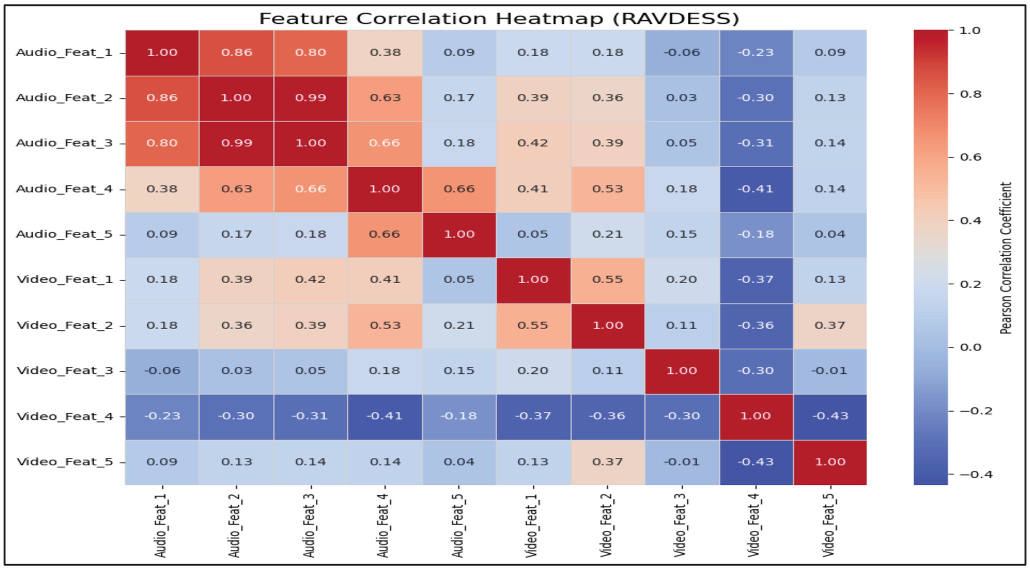

Figure 11. Feature correlation heatmap of the RAVDESS dataset.

5. Conclusions

Emotion recognition centres on the ability to precisely infer human emotions from numerous sources and modalities including physical signals such as speech and facial expression as discussed in this article. However, as recently noted by Mengara Mengara and Moon [88], “multimodal emotion recognition faces substantial challenges due to the inherent heterogeneity of data sources, each with its own temporal resolution, noise characteristics, and potential for incompleteness" [88] (p. 1). In this context, this article put forward and evaluated a novel audio-image-based cross-modal fusion model which leverages a Tucker decomposition-based factorization layer designed to explicitly capture and efficiently fuse high-order spatiotemporal feature interactions. This method reduces dimensionality while preserving complex multi-linear relationships within the fused representation, enhancing interaction expressiveness and emotion recognition accuracy. Cross-attention mechanisms are incorporated to further improve alignment, allowing the model to emphasize relevant information across modalities dynamically.

This research clearly has its limitations, in that it focuses on a limited number of modalities and just three datasets for experimentation. Nevertheless, the authors believe the methodology deployed here represents a robust model for cross-model emotion recognition by extracting modality-specific spatiotemporal features. As Praveen et al. [78] recently noted "most state-of-the- art audio-visual (A-V) fusion methods rely on recurrent networks or conventional attention mechanisms that do not effectively leverage the complementary nature of A-V modalities" [78] (p. 2486), yet the model presented here facilitates the transformation of heterogeneous inputs, originating from the same emotional expression, into a unified latent representation space. The results consistently demonstrate that the proposed fusion approach outperforms the traditional fusion approach and several state-of-the-art models, achieving superior performance in emotion recognition tasks across these diverse datasets. This opens up several promising avenues for future research. For example, incorporating additional modalities and expanding datasets to include a broader spectrum of diverse emotional expressions could lead to more comprehensive and nuanced emotion recognition models. Further exploration could focus on implementing real-time emotion recognition systems in a range of organisational contexts, including enhanced customer service, adaptive educational tools, and supportive healthcare interventions.

Author Contributions: Conceptualization, H.K. and M.A.; methodology, H.K., M.A. and M.W.; software, H.K.; validation, M.A. and M.W.; formal analysis, H.K. and M.A.; investigation, H.K. and M.A.; resources, H.K.; data curation, H.K.; writing—original draft preparation, H.K.; writing—review and editing, M.A. and M.W.; visualization, H.K., M.A. and M.W.; supervision, M.A. project administration, M.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: This study did not involve interviews or surveys, and ethical review by the IRB was deemed not necessary.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are available from the corresponding author upon reasonable request and with appropriate justification.

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

Abbreviations

The following abbreviations and notations are used in this manuscript:

$\in$	Belongs to OR is an element of
$\mathbb{R}$	Set of real numbers
$\times$	Standard multiplication for scalars
$\boldsymbol{A}$	Bold uppercase letter for a matrix (e.g., $A^{D_{audio\_per\_timestep}}$)
$\mathbf{a}$	Bold lowercase letter for a vector
$\times_n$	n-mode product (multiplication of a tensor by a matrix along mode n)
N	Batch size (Number of samples in a batch)
F_{audio_seq}	Raw audio feature sequence
F_{image_seq}	Raw image feature sequence
$D_{audio\_per\_step}$	Feature dimension of audio at each time step
$D_{image\_per\_frame}$	Feature dimension of visual (image) at each frame
T_{audio}	Number of time steps/segments in the original audio feature sequence
T_{image}	Number of frames/time steps in the original visual feature sequence
T_{common}	Common temporal length after alignment
F'_{image_seq}	Cross-attention-refined image feature sequence
F'_{audio_seq}	Cross-attention-refined audio feature sequence
X	Uppercase letter, the 4th-order cross-modal input tensor to the Tucker decomposition layer, representing spatiotemporal interactions.
G	The uppercase letter, the core tensor, represents each sample's highly compressed and fused spatiotemporal representation.
K_{time}	Reduced rank for the common temporal mode in the core tensor
K_{audio}	Reduced rank for the audio feature mode in the core tensor
K_{image}	Reduced rank for the visual feature mode in the core tensor
$A^{(T_{common})}$	Factor matrix for the common temporal mode
$A^{D_{audio\_per\_timestep}}$	Factor matrix for the audio feature mode
$A^{D_{image\_per\_frame}}$	Factor matrix for the visual feature mode

Appendix A. Model Architecture and Hyperparameter Configuration

Appendix A.1. Audio Modality Configuration

The audio modality was processed using a two-dimensional convolutional neural network (2D-CNN) architecture, specifically designed to capture joint time–frequency patterns from short audio segments for emotion recognition. The input to this model consisted of Mel-spectrograms computed from raw audio signals extracted from the video clips. During preprocessing, all audio samples were resampled to 16 kHz to ensure a consistent temporal resolution across the dataset. Each 3-s audio segment (48,000 samples) was converted into a Mel-spectrogram using a 25 ms analysis window and a 10 ms hop length, resulting in 128 Mel-frequency bands over 256 time frames. These pre-processed Mel-spectrograms were then fed directly into the 2D-CNN, enabling the network to learn hierarchical spectral–temporal representations that encode vocal tone, intensity, and rhythm, key acoustic cues for emotional expression.

Appendix A.2. Video Modality Configuration

The visual modality was pre-processed to facilitate feature extraction using the ResNet-18 architecture. Each 3-s video clip, originally recorded at 60 frames per second (fps) and comprising 180 frames, was resampled to 90 representative frames to ensure temporal uniformity and computational efficiency. Each frame was subsequently resized to 224 × 224 pixels and normalized using ImageNet's mean and standard deviation values, maintaining consistency with the ResNet-18 pre-training configuration. These pre-processed RGB frames (224 × 224 × 3) were then sequentially input into the pre-trained

ResNet-18 model, from which a 512-dimensional feature vector was extracted per frame, representing high-level spatial information. To maintain a uniform temporal dimension for downstream fusion and temporal modeling, each video's feature sequence was standardized to 90 frames, resulting in a final visual representation of 90 × 512 features for the visual encoder.

Appendix A.3. Cross-Modal Attention Mechanism

A crucial attention mechanism was employed to effectively process audio and video features. This mechanism allows the model to dynamically weigh the importance of elements from one modality when processing the other, thereby learning intricate inter-modal relationships. The architecture incorporates two distinct cross-attention modules:

1. Visual attending to Audio: The visual features sequence acts as the query, seeking relevant information from the audio sequence, which serves as the keys and values. This process generates a visual sequence enriched audio context.
2. Audio attending to Visual: Conversely, the audio feature sequence acts as the query, attending to the visual feature sequence (key and values). This results in an audio sequence enriched with visual context.

Both cross-attention modules were configured with a common embedding dimension of 256 and utilized 4 attention heads, enabling parallel learning of diverse cross-modal interactions. The output of these modules consists of context-aware feature sequences, where relevant elements from the other inform each element in one modality

As regards the temporal alignment mechanism between the audio and visual modalities, the audio and video feature sequences extracted by the 2D CNN, input size (60 × 128), and ResNet-18, input size (90 × 512) initially have different temporal resolutions due to differences in sampling rates and frame rates. To enable synchronized cross-modal fusion, both modalities are projected into a common latent dimension (256), and the visual sequence is temporally aligned to 60 steps ($T_{common} = 60$) using interpolation. This ensures that each audio frame corresponds to the same temporal segment in the visual stream.

The aligned features (audio: 60 × 256, video: 60 × 256) are then utilized in the cross-modal attention stage, enabling frame-to-frame correspondence prior to Tucker decomposition and FBP fusion. This clarification, along with the detailed dimensional flow, has been added to the revised manuscript.

Appendix A.4. The Inputs of the Proposed Architecture

The audio and visual inputs undergo modality-specific preprocessing before being fed into the fusion pipeline. The audio stream is first converted into Mel-spectrograms of size 128 × 256 (Mel bins × time frames), which serve as direct inputs to a 2D-CNN designed to learn hierarchical spectral–temporal representations capturing vocal tone, intensity, and rhythm (key acoustic cues) for emotion expression. The output of the audio encoder consists of feature embeddings of size 60 × 128, representing 60 temporal frames and 128 feature dimensions per frame. The visual stream is processed using a ResNet-18 model pretrained on ImageNet, taking input frames of size 224 × 224 × 3. Each 3-s video clip is standardized to 90 frames, with each frame encoded as a 512-dimensional feature vector, resulting in a final visual feature sequence of 90 × 512.

To facilitate multimodal fusion, linear projection layers are applied to align both modalities in a common latent space, transforming the audio and visual features to 60 × 256 and 90 × 256, respectively. A cross-modal multi-head attention mechanism (MH = 4) is then employed to model inter-modality dependencies, producing contextually attended feature sequences of dimension 60 × 256 for both modalities. The resulting attention-weighted representations are integrated into a Tucker decomposition module for spatiotemporal compression, where the input tensor of size 60 × 256 × 256 is reduced to a compact representation of 30 × 128 × 128. These compressed multimodal features are subsequently fused using a Factorized Bilinear Pooling (FBP) mechanism, generating a unified feature vector of dimension N × 512. Finally, this fused representation is passed through a Softmax classification layer, yielding the final emotion prediction output of dimension N × 7, corresponding to the seven emotion categories.

Table A1. Hyperparameters for 2D CNN and ResNet-18 model.

Audio Modality	Video Modality
2D CNN	Resnet-18
Input size = (Batch size, 128 × 256)	Input size = 224 × 224 × 3
Kernel size = 3	Kernel (conv layers) = 7 × 7, 3 × 3, 1 × 1
Activation function = ReLU	Activation function = ReLU
Max pooling = 3 × 3, stride = 2	Max pooling = 3 × 3, stride = 2
Batch size =128	Batch size = 128
Epochs = 50–100	Epochs = 50–100
Learning rate = 0.0001 to 0.001	Learning rate = 0.0001 to 0.001
Optimizer = Adam	Optimizer= Adam
Loss Function = Categorical Cross-Entropy	Loss Function = Categorical Cross-Entropy
Audio Augmentation = Pitch shifting, time stretching, background noise	Video Augmentation Horizontal flip, rotation, random cropping
Synchronization = Temporal alignment of audio and video segments	Synchronization = Temporal alignment of audio and video segments
Output dimension = 128	Output dimension = 512
Latent Space dimension = 60 × 256	Latent Space dimension = 90 × 256
Random 80:20 Split	
Number of Attention_head = 4	
Dropout_rate = 0.5	
Weight_decay= 1×10^{-4}	

Appendix B. The Input Tensor, Tensor Construction, and Tucker Decomposition for Fusion, and Associated Algorithms

Enriching the input tensor (X): Instead of constructing a 3rd-order tensor $X \in \mathbb{R}^{N \times D_{audio} \times D_{image}}$ by taking the outer product of the final, aggregated audio (F_{audio}) and image (F_{image}) feature vectors, our approach aims to preserve more granular temporal information. To achieve this, we obtain a higher-order feature sequence from each modality:

Audio features: As shown in Equation (A1), our 2D-CNN is designed to produce a feature sequence-

$$F'_{audio\_seq} \in \mathbb{R}^{T_{audio} \times D_{audio\_per\_step}} \tag{A1}$$

T_{audio} represents the number of temporal steps, and $D_{audio\_per\_step}$ is the feature dimension for each time step.

Visual features: Similarly, for ResNet-18, after processing each video frame separately and potentially refining with a GRU, we obtain a visual feature sequence as defined in Equation (A2):

$$F'_{image\_seq} \in \mathbb{R}^{T_{image} \times D_{image\_per\_frame}} \tag{A2}$$

T_{image} is the number of video frames (or time steps), and $D_{image\_per\_frame}$ is the feature dimensions per frame.

Tensor Construction: Following temporal alignment and potentially cross-modal attention, the refined audio and visual feature sequences are used to form a 3rd order tensor for each sample, as shown in Equation (A3):

$$\left(X_{sample} \in \mathbb{R}^{T_{common} \times D_{audio\_per\_timestep} \times D_{image\_per\_frame}}\right) \tag{A3}$$

Each element X_{sample} (t,d$_a$,d$_i$) explicitly captures the pairwise interaction between audio feature d$_a$ at common time-step t and visual feature d$_i$ at a common time step t. This tensor implicitly captures spatiotemporal interactions across modalities. Extending to a batch of N samples, the complete final input tensor for the Tucker decomposition is defined in Equation (A4):

$$X \in \mathbb{R}^{N\times T_{common} \times D_{audio\_per\_timestep}\times D_{image\_per\_frame}} \quad \text{(A4)}$$

Tucker decomposition for fusion: Finally, Tucker decomposition is applied to the 4th-order cross-modal tensor, X as expressed in Equation (A5):

$$X = G \times_2 A^{(T_{common})} \times_3 A^{D_{audio\_per\_timestep}} \times_4 A^{D_{image\_per\_frame}} \quad \text{(A5)}$$

$G = \mathbb{R}^{N \times K_{time}\times K_{audio}\times K_{image}}$ is the core tensor. This core tensor represents the highly compressed, fused spatiotemporal representation for each sample in the batch, where N is the original batch size.

$A^N \in \mathbb{R}^{N\times K_N}$ is the factor matrix for the batch dimension. K_N would be a compressed rank for the batch dimension ($K_N \leq N$).

$A^{(T_{common})} \in \mathbb{R}^{T_{common}\times K_{time}}$ is the factor matrix compressing the common temporal dimension from T_{common} to K_{time}.

$A^{D_{audio\_per\_timestep}} \in \mathbb{R}^{D_{audio\_per\_timestep}\times K_{audio}}$ is the matrix factor that compresses the audio feature dimension per time step from $D_{audio\_per\_timestep}$ to K_{audio}.

$A^{D_{image\_per\_frame}} \in \mathbb{R}^{D_{image\_per\_frame}\times K_{image}}$ is the factor matrix that compresses the visual feature dimension per frame from $D_{image\_per\_frame}$ to K_{image}.

The pruning ranks in the Tucker decomposition were selected manually based on empirical evaluation and dimensionality constraints of the multimodal feature space. The temporal and feature-mode ranks were chosen to provide sufficient expressive power while reducing the original tensor dimensions (60 × 256 × 256) to a more compact form (30 × 128 × 128). These fixed ranks are used throughout training, with the factor matrices learned jointly with the rest of the network.

Algorithm A1: Tensor construction from audio and visual features

```
Input:
  •  Attention-weighted audio feature sequence for sample n
       $\left(F'_{audio_{seq},n} \in \mathbb{R}^{T_{common} \times D_{audio\_per\_timestep}}\right)$
  •  Attention-weighted visual feature sequence for sample n
       $\left(F'_{image_{seq},n} \in \mathbb{R}^{T_{common} \times D_{image\_per\_frame}}\right)$
Output:
  •  A 4th-order cross-modal tensor
       $X \in \mathbb{R}^{N\times T_{common} \times D_{audio\_per\_timestep}\times D_{image\_per\_frame}}$
Procedure:
  1. Initialize Tensor List = [ ]
  2. For each sample n = 1 to N:
     a. Initialize $X_{sample_n} \in \mathbb{R}^{T_{common} \times D_{audio\_per\_timestep}\times D_{image\_per\_frame}}$
     b. For each time step t-=1 to $T_{common}$ :
        i.   Extract audio feature vector:
               $a_t = F'_{audio\_seq_n}(t) \in \mathbb{R}^{D_{audio\_per\_timestep}}$
        ii.  Extract visual feature vector:
               $v_t = F'_{image\_seq_n}(t) \in \mathbb{R}^{D_{image\_per\_frame}}$
        iii. Compute the outer product:
               $M_t = a_t \otimes v_t \in \mathbb{R}^{D_{audio\_per\_timestep}\times D_{image\_per\_frame}}$
        iv.  Store as the tth slice: $X_{sample}(t,:,:) = M_t$
     c. Append $X_{sample}$ to Tensor List
  3. Stack all sample tensors along a new first dimension:
       $X = Stack\ (Tensor\_List) \in \mathbb{R}^{N\times T_{common} \times D_{audio\_per\_timestep}\times D_{image\_per\_frame}}$
```

Algorithm A2: Cross-modal spatiotemporal fusion by Tucker decomposition layer

```
Input:
    • A 4th-order cross-modal tensor
        $X \in \mathbb{R}^{N \times T_{common} \times D_{audio\_per\_timestep} \times D_{image\_per\_frame}})$
    • Desired output ranks for the core tensor: $K_{time} \times K_{audio} \times K_{image}$
Output:
    • The fused low-rank core tensor $G = \mathbb{R}^{N \times K_{time} \times K_{audio} \times K_{image}}$
Learnable parameters (End-to-End Trainable):
    • Factor matrix for common temporal dimension:
        $A^{(T_{common})} \in \mathbb{R}^{T_{common} \times K_{time}}$
    • Factor matrix for audio feature dimension:
        $A^{D_{audio\_per\_timestep}} \in \mathbb{R}^{D_{audio\_per\_timestep} \times K_{audio}}$
    • Factor matrix for visual feature dimension:
        $A^{D_{image\_per\_frame}} \in \mathbb{R}^{D_{image\_per\_frame} \times K_{image}}$
Procedure:
    1. Initialization
        • InitializationInitialize factor matrices $A^{(T_{common})},\ A^{(D_{audio\_per\_timestep})},\ A^{(D_{image\_per\_frame})}$
    2. Tensor Projection via n-mode products:
        a. Mode-2 (Temporal) Projection: $X_1 \leftarrow X \times_2 (A^{(T^{common})}))^T$
           Where $X_1 \in \mathbb{R}^{N \times K_{time} \times D_{audio\_per\_timestep} \times D_{image\_per\_frame}}$
        b. Mode-3 (Audio) Projection: $X_2 \leftarrow X_1 \times_3 (A\ A^{(D_{audio\_per\_timestep})})^T$
           Where $X_2 \in \mathbb{R}^{N \times K_{time} \times K_{audio} \times D_{image\_per\_frame}}$
        c. Mode-4 (Visual) Projection: $G \leftarrow X_2 \times_4 (A^{A^{(D_{image\_per\_frame})}})^T$
           Where G $\in \mathbb{R}^{N \times K_{time} \times K_{audio} \times K_{image}}$
3. Return G as the fused representation.
```

Appendix C. Datasets Used in the Experiments

- IEMOCAP: The IEMOCAP dataset is one of the challenging and complex benchmark datasets for cross-modal emotion recognition. It provides an extensive multimodal resource for recognizing emotions and capturing emotive speech, corresponding facial expressions, motion captures, and body gestures. The dataset is annotated with multiple emotional categories, making it highly suitable for supervised learning tasks. The dataset comprises 302 video utterances (utterances in the video-Audio dictionary: 302) and consists of dyadic sessions where actors perform improvisations or scripted scenarios, captured through audio and video modalities. Each utterance is annotated with nine distinct emotion categories, enabling fine-grained emotional analysis. This dataset is pivotal for developing and evaluating emotion recognition systems, especially in cross-modal contexts.
- RAVDESS: The proposed research extended its experimental validation to include the RAVDESS dataset, which serves as a challenging and benchmark dataset for cross-modal emotion recognition. This dataset consists of video recordings from 12 male and 12 female professionals (24 actors), each contributing 60 unique samples and eight distinct emotion states. Each utterance is in .mp4 format, making the RAVDESS dataset well-suited for evaluating cross-modal fusion techniques.
- REMA-D: This dataset comprises 7442 recordings from 91 different actors. The CREMA-D dataset offers a valuable collection of spontaneous and posed emotional expressions. It covers six core emotions, with each utterance captured in both audio and video (.wav and .flv) formats. The CREMA-D was validated using crowd-sourced human ratings, providing reliable emotion labels. This dataset serves as a strong foundation for training and testing emotion recognition models across diverse speakers, modalities, and expression styles.

Table A2. Dataset description.

Dataset	Total Utterances	Modality	Emotion Labels
IEMOCAP	302	Audio-Video	Neutral, Calm, Happy, Sad, Angry, Fear, Disgust, Excited, Surprised
RAVDESS	2857	Audio-Video	Neutral, Calm, Happy, Sad, Angry, Fear, Disgust, Surprise
CREMA-D	7442	Audio-Video	Neutral, Happy, Sad, Angry, Fearful, Disgust

Appendix D. System Configurations

All experiments were conducted on a high-performance system configured with an NVIDIA GeForce RTX 3090 GPU, an Intel Core i9 processor, and 128 GB of RAM, ensuring sufficient computational capacity for training deep learning models on large-scale cross-modal datasets. The software environment was based on the Ubuntu 20.04 LTS operating system, with Pytorch 1.13.0 serving as the primary deep-learning framework for model development and experimentation. All implementations were carried out using Python 3.11.1 programming language within the Jupyter Notebook integrated development environment (IDE).

References

1. Cohn, J.F.; Ambadar, Z.; Ekman, P. Observer-based measurement of facial expression with the Facial Action Coding System. In *Handbook of Emotion Elicitation and Assessment*; Coan, J.A., Allen, J.J.B., Eds.; Oxford University Press: Oxford, UK, 2007; pp. 203–221.
2. Avital, N.; Egel, I.; Weinstock, I.; Malka, D. Enhancing Real-Time Emotion Recognition in Classroom Environments Using Convolutional Neural Networks: A Step Towards Optical Neural Networks for Advanced Data Processing. *Inventions* **2024**, *9*, 113. [CrossRef]
3. Kalateh, S.; Estrada-Jimenez, L.A.; Nikghadam-Hojjati, S.; Barata, J. A Systematic Review on Multimodal Emotion Recognition: Building Blocks, Current State, Applications, and Challenges. *IEEE Access* **2024**, *12*, 103976–104019. [CrossRef]
4. Hu, P.; Huang, Y.; Mei, J.; Leung, H.; Chen, Z.; Kuang, Z.; You, Z.; Hu, L. Learning from low-rank multimodal representations for predicting disease-drug associations. *BMC Med. Inf. Decis. Mak.* **2021**, *21*, 308. [CrossRef]
5. DeVault, D.; Artstein, R.; Benn, G.; Dey, T.; Fast, E.; Gainer, A.; Georgila, K.; Gratch, J.; Hartholt, A.; Lhommet, M.; et al. SimSensei kiosk: A virtual human interviewer for healthcare decision support. In Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems, Paris, France, 5–9 May 2014; Volume 2, pp. 1061–1068.
6. Khan, W.A.; Qudous, H.; Farhan, A.A. Speech emotion recognition using feature fusion: A hybrid approach to deep learning. *Multimed. Tools Appl.* **2024**, *83*, 75557–75584. [CrossRef]
7. Zhou, F.; Kong, S.; Fowlkes, C.C.; Chen, T.; Lei, B. Fine-grained facial expression analysis using dimensional emotion model. *Neurocomputing* **2020**, *392*, 38–49. [CrossRef]
8. Kuchibhotla, S.; Yalamanchili, B.S.; Vankayalapati, H.D.; Anne, K.R. Speech Emotion Recognition Using Regularized Discriminant Analysis. *Adv. Intell. Syst. Comput.* **2014**, *247*, 363–369. [CrossRef]
9. Ortega, J.D.S.; Cardinal, P.; Koerich, A. Emotion recognition using fusion of audio and video features. In Proceedings of the 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC), Bari, Italy, 6–9 October 2019; pp. 3847–3852. [CrossRef]
10. Raj, R.S.; Pratiba, D.; Kumar, R.P. Facial Expression Recognition using Facial Landmarks: A novel approach. *Adv. Sci. Technol. Eng. Syst.* **2020**, *5*, 24–28. [CrossRef]
11. Ristea, N.C.; Dutu, L.C.; Radoi, A. Emotion recognition system from speech and visual information based on convolutional neural networks. In Proceedings of the 2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), Timisoara, Romania, 10–12 October 2019; pp. 1–6. [CrossRef]
12. Chaudhari, A.; Bhatt, C.; Nguyen, T.T.; Patel, N.; Chavda, K.; Sarda, K. Emotion Recognition System via Facial Expressions and Speech Using Machine Learning and Deep Learning Techniques. *SN Comput. Sci.* **2023**, *4*, 363. [CrossRef]
13. Rangulov, D.; Fahim, M. Emotion Recognition on large video dataset based on Convolutional Feature Extractor and Recurrent Neural Network. In Proceedings of the International Conference on Image Processing, Applications and Systems, IPAS, Genoa, Italy, 9–11 December 2020; pp. 14–20.
14. Tholusuri, A.; Anumala, M.; Malapolu, B.; Jaya Lakshmi, G. Sentiment analysis using LSTM. *Int. J. Eng. Adv. Technol.* **2019**, *8*, 1338–1340. [CrossRef]
15. Liu, K.; Feng, Y.; Zhang, L.; Wang, R.; Wang, W.; Yuan, X.; Cui, X.; Li, X.; Li, H. An Effective Personality-Based Model for Short Text Sentiment Classification Using BiLSTM and Self-Attention. *Electron* **2023**, *12*, 3274. [CrossRef]

16. Pan, X.; Guo, W.; Guo, X.; Li, W.; Xu, J.; Wu, J. Deep temporal-spatial aggregation for video-based facial expression recognition. *Symmetry* **2019**, *11*, 52. [CrossRef]
17. Sanku, S.R.; Sandhya, B. Multi-Modal Emotion Recognition Feature Extraction and Data Fusion Methods Evaluation. *Int. J. Innov. Technol. Explor. Eng.* **2024**, *3075*, 18–27. [CrossRef]
18. Zhang, K.; Li, Y.; Wang, J.; Wang, Z.; Li, X. Feature fusion for multimodal emotion recognition based on deep canonical correlation analysis. *IEEE Signal Process. Lett.* **2021**, *28*, 1898–1902. [CrossRef]
19. Hazarika, D.; Gorantla, S.; Poria, S.; Zimmermann, R. Self-Attentive Feature-level Fusion for Multimodal Emotion Detection. In Proceedings of the 2018 IEEE Conference on multimedia information processing and retrieval (MIPR), Miami, FL, USA, 10–12 April 2018; pp. 196–201. [CrossRef]
20. Dixit, C.; Satapathy, S.M. Deep CNN with late fusion for real time multimodal emotion recognition. *Expert Syst. Appl.* **2024**, *240*, 122579. [CrossRef]
21. Kumar, P.; Malik, S.; Raman, B. Interpretable multimodal emotion recognition using hybrid fusion of speech and image data. *Multimed Tools Appl.* **2024**, *83*, 28373–28394. [CrossRef]
22. Gill, J.; Johnson, P. *Research Methods for Managers*, 3rd ed.; Sage: London, UK, 2002.
23. Porter, A.L.; Kongthon, A.; Lu, J.C. Research Profiling: Improving the Literature Review. *Scientometrics* **2002**, *53*, 351–370. [CrossRef]
24. Popay, J.; Roberts, H.; Sowden, A.; Petticrew, M.; Arai, L.; Rodgers, M. *Guidance on the Conduct of Narrative Synthesis in Systematic Reviews*; Lancaster University: Lancaster, UK, 2006.
25. Arksey, H.; O'Malley, L. Scoping studies: Towards a methodological framework. *Int. J. Soc. Res. Methodol.* **2005**, *8*, 19–32. [CrossRef]
26. Snyder, H. Literature review as a research methodology: An overview and guidelines. *J. Bus. Res.* **2019**, *104*, 333–339. [CrossRef]
27. Levering, B. Concept Analysis as Empirical Method. *Int. J. Qual. Methods* **2002**, *1*, 35–48. [CrossRef]
28. Jabareen, Y. Building a Conceptual Framework: Philosophy, Definitions, and Procedure. *Int. J. Qual. Methods* **2009**, *8*, 49–62. [CrossRef]
29. Indeed Editorial Team. Experimental Research: Definition, Types and Examples. 2024. Available online: https://www.indeed.com/career-advice/career-development/experimental-research (accessed on 10 October 2025).
30. Zadeh, A.; Chen, M.; Cambria, E.; Poria, S.; Morency, L.P. Tensor fusion network for multimodal sentiment analysis. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 7–11 September 2017; pp. 1103–1114. [CrossRef]
31. Tucker, L.R. Some mathematical notes on three-mode factor analysis. *Psychometrika* **1966**, *31*, 279–311. [CrossRef] [PubMed]
32. Ben-younes, H.; Cord, M.; Thome, N. MUTAN: Multimodal Tucker Fusion for Visual Question Answering. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2612–2620. [CrossRef]
33. Mai, S.; Hu, H.; Xing, S. Modality to Modality Translation: An Adversarial Representation Learning and Graph Fusion Network for Multimodal Fusion. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 164–172.
34. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
35. Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; Fei-fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255. [CrossRef]
36. Wang, R.; Zhu, J.; Wang, S.; Wang, T.; Huang, J.; Zhu, X. Multi-modal emotion recognition using tensor decomposition fusion and self-supervised multi-tasking. *Int. J. Multimed. Inf. Retr.* **2024**, *13*, 39. [CrossRef]
37. Ben-younes, H.; Cadene, R.; Thome, N.; Cord, M. BLOCK: Bilinear Superdiagonal Fusion for Visual Question Answering and Visual Relationship Detection. In Proceedings of the Third AAAI Conference on Artificial Intelligence, Washington, DC, USA, 22–26 August 2019; pp. 8102–8109. [CrossRef]
38. Busso, C.; Bulut, M.; Lee, C.C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J.; Lee, S.; Narayanan, S. IEMOCAP: Interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* **2008**, *42*, 335–359. [CrossRef]
39. Livingstone, S.R.; Russo, F.A. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLoS ONE* **2018**, *13*, e0196391. [CrossRef]
40. Houwei, C.; Cooper, D.; Keutmann, M.; Gur, R.; Nenkova, A.; Ragini, V. CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset. *IEEE Trans. Affect. Comput.* **2014**, *5*, 377–390. [CrossRef]
41. Murugappan, M.; Mutawa, A. Facial geometric feature extraction based emotional expression classification using machine learning algorithms. *PLoS ONE* **2021**, *16*, e0247131. [CrossRef]
42. Roopa, N.S.; Prabhakaran, M.; Betty, P. Speech emotion recognition using deep learning. *Int. J. Recent Technol. Eng.* **2019**, *7*, 247–250. [CrossRef]

43. Akhand, M.A.H.; Roy, S.; Siddique, N.; Kamal, M.A.S.; Shimamura, T. Facial emotion recognition using transfer learning in the deep CNN. *Electronics* **2021**, *10*, 1036. [CrossRef]
44. Liu, G.; Cai, S.; Wang, C. Speech emotion recognition based on emotion perception. *Eurasip J. Audio Speech Music. Process.* **2023**, *1*, 22. [CrossRef]
45. 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]
46. Sultana, T.; Jahan, M.; Uddin, K.; Kobayashi, Y.; Smieee, M.H. Multimodal Emotion Recognition through Deep Fusion of Audio-Visual Data. In Proceedings of the 26th International Conference on Computer and Information Technology (ICCIT), Cox's Bazar, Bangladesh, 13–15 December 2023; IEEE: Bissen, Luxembourg, 2023; pp. 1–5. [CrossRef]
47. Goncalves, L.; Leem, S.G.; Lin, W.C.; Sisman, B.; Busso, C. Versatile Audio-Visual Learning for Emotion Recognition. *IEEE Trans. Affect. Comput.* **2024**, *16*, 306–318. [CrossRef]
48. John, V.; Kawanishi, Y. Audio and Video-based Emotion Recognition using Multimodal Transformers. In Proceedings of the International Conference on Pattern Recognition (ICPR), Montreal, QC, Canada, 21–25 August 2022; pp. 2582–2588. [CrossRef]
49. Singh, N.; Khan, R.; Shree, R. MFCC and Prosodic Feature Extraction Techniques: A Comparative Study. *Int. J. Comput. Appl.* **2012**, *54*, 9–13. [CrossRef]
50. Aouani, H.; Ayed, Y.B. Speech Emotion Recognition with deep learning. In Proceedings of the 24th International Conference on Knowledge-Based and Intelligent Information & Engineering Speech Emotion Recognition with Deep Learning Systems, Verona, Italy, 16–18 September 2020; Elsevier: Amsterdam, The Netherlands, 2020; pp. 251–260.
51. Sharma, G.; Umapathy, K.; Krishnan, S. Trends in audio signal feature extraction methods. *Appl. Acoust.* **2020**, *158*, 20. [CrossRef]
52. Mehrish, A.; Majumder, N.; Bharadwaj, R.; Mihalcea, R.; Poria, S. A review of deep learning techniques for speech processing. *Inf. Fusion* **2023**, *99*, 101869. [CrossRef]
53. Lim, W.; Jang, D.; Lee, T. Speech Emotion Recognition using Convolutional Recurrent Neural Networks and Spectrograms. In Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), Jeju, Republic of Korea, 13–16 December 2016; IEEE: Bissen, Luxembourg, 2016; pp. 1–5. [CrossRef]
54. Li, C.; Bao, Z.; Li, L.; Zhao, Z. Exploring temporal representations by leveraging attention-based bidirectional LSTM-RNNs for multi-modal emotion recognition. *Inf. Process Manag.* **2020**, *57*, 102185. [CrossRef]
55. Venkateswarlu, S.C.; Jeevakala, S.R.; Kumar, N.U.; Munaswamy, P.; Pendyala, D. Emotion Recognition From Speech and Text using Long Short-Term Memory. *Eng. Technol. Appl. Sci. Res.* **2023**, *13*, 11166–11169. [CrossRef]
56. Shaikh, M.B.; Chai, D.; Islam, S.M.S.; Akhtar, N. Multimodal fusion for audio-image and video action recognition. *Neural. Comput. Appl.* **2024**, *5*, 5499–5513. [CrossRef]
57. Palash, M.; Bhargava, B. EMERSK-Explainable Multimodal Emotion Recognition with Situational Knowledge. *IEEE Trans. Multimed.* **2023**, *26*, 2785–2794. [CrossRef]
58. Zhang, S.; Tao, X.; Chuang, Y.; Zhao, X. Learning deep multimodal affective features for spontaneous speech emotion recognition. *Speech Commun.* **2021**, *127*, 73–81. [CrossRef]
59. Lakshmi, K.L.; Muthulakshmi, P.; Nithya, A.A.; Jeyavathana, R.B.; Usharani, R.; Das, N.S.; Devi, G.N.R. Recognition of emotions in speech using deep CNN and RESNET. *Soft. Comput.* **2023**, 1–16. [CrossRef]
60. Udeh, C.P.; Chen, L.; Du, S.; Li, M.; Wu, M. Multimodal Facial Emotion Recognition Using Improved Convolution Neural Networks Model. *J. Adv. Comput. Intell. Intell. Inform.* **2023**, *27*, 710–719. [CrossRef]
61. Patil, G.; Suja, P. Emotion Recognition from 3D Videos using Optical Flow Method. In Proceedings of the International Conference on Smart Technology for Smart Nation, (SmartTechCon), Bangalore, India, 17–19 August 2017; IEEE: Bissen, Luxembourg, 2017; pp. 825–829.
62. Kumari, N.; Bhatia, R. Deep learning based efficient emotion recognition technique for facial images. *Int. J. Syst. Assur. Eng. Manag.* **2023**, *14*, 1421–1436. [CrossRef]
63. Alkawaz, M.H.; Mohamad, D.; Basori, A.H.; Saba, T. Blend Shape Interpolation and FACS for Realistic Avatar. *3D Res.* **2015**, *6*, 6. [CrossRef]
64. Cho, J.; Hwang, H. Spatio-temporal representation of an electoencephalogram for emotion recognition using a three-dimensional convolutional neural network. *Sensors* **2020**, *20*, 3491. [CrossRef] [PubMed]
65. Adegun, I.P.; Vadapalli, H.B. Facial micro-expression recognition: A machine learning approach. *Sci. Afr.* **2020**, *8*, 14. [CrossRef]
66. Mehta, N.; Jadhav, S. Facial Emotion recognition using Log Gabor filter and PCA Ms Neelum Mehta. In Proceedings of the International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 12–13 August 2016; IEEE: Bissen, Luxembourg, 2016; pp. 1–5.
67. Shi, Y.; Lv, Z.; Bi, N.; Zhang, C. An improved SIFT algorithm for robust emotion recognition under various face poses and illuminations. *Neural. Comput. Appl.* **2020**, *32*, 9267–9281. [CrossRef]
68. Lakshmi, D.; Ponnusamy, R. Facial emotion recognition using modified HOG and LBP features with deep stacked autoencoders. *Microprocess. Microsyst.* **2021**, *82*, 103834. [CrossRef]

69. Schoneveld, L.; Othmani, A.; Abdelkawy, H. Leveraging recent advances in deep learning for audio-Visual emotion recognition. *Pattern Recognit. Lett.* **2021**, *146*, 1–7. [CrossRef]
70. Yu, C.; Zhao, X.; Zheng, Q.; Zhang, P.; You, X. Hierarchical Bilinear Pooling for Fine-Grained Visual Recognition. In Proceedings of the 15th European Conference on computer vision (ECCV), Munich, Germany, 8–14 September 2018; Volume 11220, pp. 595–610. [CrossRef]
71. Sahoo, S.; Routray, A. Emotion recognition from audio-visual data using rule based decision level fusion. In Proceedings of the IEEE Students' Technol Symp TechSym 2016, Kharagpur, India, 2 October 2016; pp. 7–12. [CrossRef]
72. Shoumy, N.J.; Ang, L.M.; Seng, K.P.; Rahaman, D.M.M.; Zia, T. Multimodal big data affective analytics: A comprehensive survey using text, audio, visual and physiological signals. *J. Netw. Comput. Appl.* **2020**, *149*, 102447. [CrossRef]
73. Ortega, J.D.S.; Senoussaoui, M.; Granger, E.; Pedersoli, M.; Cardinal, P.; Koerich, A.L. Multimodal Fusion with Deep Neural Networks for Audio-Video Emotion Recognition. *arXiv* **2019**, arXiv:190703196. [CrossRef]
74. Njoku, J.N.; Caliwag, A.C.; Lim, W.; Kim, S.; Hwang, H.J.; Jeong, J.W. Deep Learning Based Data Fusion Methods for Multimodal Emotion Recognition. *J. Korean Inst. Commun. Inf. Sci.* **2022**, *47*, 79–87. [CrossRef]
75. Cimtay, Y.; Ekmekcioglu, E.; Caglar-Ozhan, S. Cross-subject multimodal emotion recognition based on hybrid fusion. *IEEE Access* **2020**, *8*, 168865–168878. [CrossRef]
76. Yoshino, K.; Sakti, S.; Nakamura, S. Hierarchical Tensor Fusion Network for Deception Handling Negotiation Dialog Model. In Proceedings of the 33rd Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; pp. 1–10. Available online: https://neurips.cc/virtual/2019/workshop/13200 (accessed on 10 October 2025).
77. Krishna, D.N.; 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.
78. 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 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), New Orleans, LA, USA, 19–20 June 2022; pp. 2485–2494. [CrossRef]
79. Zhou, S.; Wu, X.; Jiang, F.; Huang, Q.; Huang, C. Emotion Recognition from Large-Scale Video Clips with Cross-Attention and Hybrid Feature Weighting Neural Networks. *Int. J. Environ. Res. Public Health* **2023**, *20*, 1400. [CrossRef]
80. Lee, Y.; Yoon, S.; Jung, K. Multimodal Speech Emotion Recognition using Cross Attention with Aligned Audio and Text. In Proceedings of the Interspeech, Shanghai, China, 25–29 October 2020; ISCA: Geneva, Switzerland, 2020; pp. 2717–2721.
81. Liu, F.; Fu, Z.; Wang, Y.; Zheng, Q. TACFN: Transformer-Based Adaptive Cross-Modal Fusion Network for Multimodal Emotion Recognition. *CAAI Artif. Intell. Res.* **2023**, *2*, 9150019. [CrossRef]
82. Fu, Z.; Liu, F.; Wang, H.; Qi, J.; Fu, X.; Zhou, A.; Li, Z. A cross-modal fusion network based on self-attention and residual structure for multimodal emotion recognition. *arXiv* **2021**, arXiv:2111.02172. [CrossRef]
83. Luna-Jiménez, C.; Kleinlein, R.; Griol, D.; Callejas, Z.; Montero, J.M.; Fernández-Martínez, F. A proposal for multimodal emotion recognition using aural transformers and action units on ravdess dataset. *Appl. Sci.* **2022**, *12*, 327. [CrossRef]
84. Jin, Z.; Zai, W. Audiovisual emotion recognition based on bi-layer LSTM and multi-head attention mechanism on RAVDESS dataset. *J. Supercomput.* **2025**, *81*, 31. [CrossRef]
85. Moorthy, S.; Moon, Y.K. Hybrid Multi-Attention Network for Audio–Visual Emotion Recognition Through Multimodal Feature Fusion. *Mathematics* **2025**, *13*, 1100. [CrossRef]
86. Feng, J.; Fan, X. Cross-modal Context Fusion and Adaptive Graph Convolutional Network for Multimodal Conversational Emotion Recognition. *arXiv* **2025**, arXiv:2501.15063. [CrossRef]
87. Hu, D.; Chen, C.; Zhang, P.; Li, J.; Yan, Y.; Zhao, Q. A two-stage attention based modality fusion framework for multi-modal speech emotion recognition. *IEICE Trans. Inf. Syst.* **2021**, *E104D*, 1391–1394. [CrossRef]
88. Mengara Mengara, A.G.; Moon, Y.K. CAG-MoE: Multimodal Emotion Recognition with Cross-Attention Gated Mixture of Experts. *Mathematics* **2025**, *13*, 1907. [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.