ATTSF-Net: Attention-Based Similarity Fusion Network for Audio-Visual Emotion Recognition

Jiaming Zhang, Zhijia Zhang, and Zhaojie Ju, *Senior Member, IEEE*

***Abstract*—Emotional factors play a pivotal role in fields such as autonomous driving and intelligent emotional robotics. The accurate extraction of emotional factors is instrumental in reducing error rates within these domains. With the continuous deepening of exploration in the emotional domain, rich multimodal data has progressively supplanted unimodal data. Nevertheless, current multimodal approaches still grapple with the following challenges: 1) Partial loss of information both within individual modalities and across different modalities. 2) Incorrect extraction of modality-invariant features. To facilitate multimodal interaction and address the aforementioned issues, this paper proposes an Attention-based Similarity Fusion Network (ATTSF-Net) for audio-visual emotion recognition. The network is based on multimodal data and comprises the proposed Cross-Multimodal Block (CMB), Similarity Adjustment Block (SAB), and Audio-Visual Auxiliary Modules (AVAM). CMB employs a cross-modal attention mechanism and model-level fusion to facilitate interactions between modalities. SAB is designed to learn modality-invariant features. AVAM utilizes additional audio-visual auxiliary networks to provide supplementary emotional information, enabling the full extraction of intra-modal information. A similarity loss function based on Kullback-Leibler (KL) divergence is designed to ensure the consistency of the learned audio-visual emotional information. The proposed model achieves an accuracy of 88.67% on the RAVDESS dataset (8.67% higher than human) and an unweighted accuracy (UA) of 81.93% and a weighted accuracy (WA) of 79.77% on the IEMOCAP dataset (4.01% and 5.86% higher than transfer learning, respectively). A series of visualization experiments are conducted to demonstrate the effectiveness of the proposed model.**

***Index Terms*—Attention-based similarity fusion network, audio-visual emotion recognition, modal-invariant features, auxiliary network.**

I. Introduction

EMOTION plays a crucial role in the field of HCI, with extensive application scenarios such as autonomous driving and intelligent emotional robots. Traditional emotion recognition approaches analyze individual modalities, such as audio or video. Furthermore, with the deepening exploration of the emotion domain and the rapid development of deep learning technologies, an increasing number of modalities are being utilized for emotion recognition analysis, including facial expressions [1], hand gestures, speech signals [2], eyebrow movements, and electroencephalograms (EEGs) [3]. A wealth of literature indicates that multimodal signals significantly outperform single-modal signals in terms of emotion recognition accuracy [4], [5], [6], [7]. Consequently, the fusion of different signal modalities has gradually become the main method of emotion recognition. One of the primary challenges in multimodal data lies in determining the number of modalities to use. Too few modalities may fail to adequately extract relevant emotional information, while an excessive number can lead to data redundancy. In multimodal emotion recognition, audio and video represent the most natural, ubiquitous, and information-rich modalities in human-human interactions [8]. They also encapsulate the primary information related to emotions and are the easiest to discern an individual's emotional state and intentions. Therefore, this paper will focus on processing audio and video modalities to achieve emotion recognition based on audiovisual data.

However, the current multimodal domain still grapples with the following issues: 1) Partial absence of information within and across modalities. While most models strive to thoroughly extract inter-modal information [6], [9], [10], this often comes at the expense of intra-modal information, resulting in incomplete feature learning. 2) Incorrect extraction of modality-invariant features and the issue of modality bias. Modality-invariant features reflect the inherent characteristics of modal data. Nevertheless, mere coupling between modalities may result in one modality exhibiting a tendency to lean towards another, thereby leading to the extraction of erroneous information.

Given that emotional changes are a continuous and dynamic process over time, the attention mechanism, with its capability to capture long-range dependencies within sequences, inherently holds significant advantages in addressing the aforementioned issues and extracting interactive features from multimodal data. The Transformer [11], leveraging its unique attention mechanism, has achieved remarkable success across multiple domains, including Natural Language Processing (NLP), Acoustic Scene Classification (ASC), and Computer Vision (CV). To translate the advantages of the Transformer to the domain of emotion recognition, an increasing number of researchers are concentrating on refining the attention mechanism to enhance the accuracy of emotion recognition [2], [10], [12]. In the context of multimodal fusion, a pivotal aspect involves how to fully extract both intra-modal and inter-modal information. Intra-modal

Received 14 January 2025; revised 15 July 2025; accepted 17 July 2025. Date of publication 22 July 2025; date of current version 3 December 2025. Recommended for acceptance by D. S. Johnson. *(Corresponding author: Zhijia Zhang.)*

Jiaming Zhang and Zhijia Zhang are with the School of Artificial Intelligence, Shenyang University of Technology, Shenyang 110178, China (e-mail: lintonao@163.com; zzjsut@126.com).

Zhaojie Ju is with the College of Biomedical Engineering and Instrument Science, Zhejiang University, Zhejiang 310027, China (e-mail: juzhaojie@icloud.com).

Digital Object Identifier 10.1109/TAFFC.2025.3591567

information refers to the thorough extraction of information from within a single modality, whereas inter-modal information relates to the effective interaction among modalities. Most literature tends to focus on either intra-modal or inter-modal considerations, and the optimal integration of both types of information remains an unresolved challenge [13]. The emergence of the cross-modal Transformer architecture presents a fresh perspective for multimodal fusion, leveraging the attention mechanism to promote information exchange across modalities [14].

Although the Transformer architecture demonstrates strong performance, it struggles with modal alignment for audio-video data [15]. Audio and video signals frequently contain redundant non-emotional features within their respective domains, which obstruct the model's capacity to learn the emotional features linked to audio-video interactions. Among existing methods, relying solely on cross-modal attention mechanisms without aligning audio and video semantics may fail to guarantee that the model captures the correct corresponding emotional information [15], [16], [17]. To tackle these issues, most literature introduces additional loss functions to ensure that the model precisely extracts relevant information pertaining to emotional cues [1], [18], [19], [20]. In real-world scenarios of emotion recognition, audio and video modalities inevitably suffer from noise and modality absence, making robustness a crucial requirement for emotion recognition systems. The challenge lies in performing emotion recognition when information from a single modality cannot be fully considered [6].

To achieve multimodal interaction and address the aforementioned issues, we propose CMB to facilitate inter-modal communication. However, relying solely on coupling operations between modalities may lead to problems such as incorrect extraction of modality-invariant features and modality bias, resulting in confused emotional feature representation. Therefore, we further design SAB to learn modality-invariant features and introduce a similarity loss function based on KL divergence for feature alignment [21], aiming to alleviate the aforementioned issues. This ensures that the model can fully extract emotional information between modalities. To avoid the omission of intra-modal emotional information, we additionally design AVAM to supplement the missing information. Through the integration of these components, the proposed ATTSF-Net is capable of fully extracting both inter-modal and intra-modal emotional information while correctly extracting modality-invariant features. The model has undergone extensive experimental validation on the RAVDESS [22] and IEMOCAP [23] datasets.

The specific contributions of this paper can be summarized as follows:

- This paper introduces the ATTSF-Net for multimodal emotion recognition, in which the CMB module based on cross-modal attention mechanism is proposed. The CMB module facilitates cross-modal interaction between audio and video, thereby enhancing the completeness of capturing the interrelationships within audio-visual signals.
- This paper proposes the SAB to learn modality-invariant features for multimodal information fusion. Additionally, a similarity loss function based on KL divergence is employed for feature alignment, enabling the adequate extraction of representative emotional information and addressing the modality bias issue inherent in traditional cross-modal models.
- Considering the issue of intra-modal information omission in traditional methods, this paper proposes the AVAM based on multi-head attention to ensure the comprehensive extraction of intra-modal information. This, in turn, enhances the accuracy and robustness of emotion recognition.
- The proposed model achieves an accuracy of 88.67% on the RAVDESS dataset and 81.93% UA and 79.77% WA on the IEMOCAP dataset. Additionally, robustness experiments are conducted, and a series of visualizations are provided to validate the accuracy of the model.

The structure of this paper is as follows: Section II presents related work in the field of emotion recognition. Section III provides a detailed introduction to the proposed model along with the corresponding details. Section IV describes the datasets and the specifics of model training. Section V outlines the experiments conducted. Finally, Section VI presents the conclusions and future prospects.

II. Related Works

Emotion recognition based on audio-video fusion involves analyzing audio and video modalities, utilizing a modality fusion network to process intra- and inter-modal relevant information, providing a comprehensive understanding of emotional states for HCI and affective computing, and enhancing the accuracy of emotion recognition. Multimodal fusion networks can be broadly categorized into three types according to their fusion methods: feature-level fusion [9], [24], decision-level fusion [5], [18], [25], [26], and model-level fusion [2], [6], [27], [28], [29].

In feature-level fusion, features extracted from multiple modalities are concatenated into a unified vector for emotion recognition. Praveen et al. [9] introduce the Audio Gating Layer, which utilizes a simple fully connected layer to calculate the weight distribution of different modalities based on the emotional information they contain. It also considers solutions for situations where one modality may be problematic. While feature-level fusion is relatively straightforward to implement, this approach only takes into account the relationships between modalities and neglects the differences in emotional features across different modalities. This strategy struggles to establish temporal synchronization between modalities.

For decision-level fusion, separate analyses are conducted on different modalities, and only the posterior probabilities from two individual classifiers are combined, using methods such as weighted combination or support vector machines, to obtain the final recognition result. Tellamekala et al. [18] propose the Uncertainty-Aware Audio-Visual Context fusion method, which treats typical deterministic embeddings as multivariate normal distributions and utilizes variances to assign weights to corresponding modalities. Additionally, Calibrated and Ordinal Latent Distributions constraints are introduced, with constraints like the Calibrated Constraint used to limit the variance of the corresponding distributions. Atmaja et al. [26] employ both feature-level and decision-level fusion, utilizing Long

Short-Term Memory (LSTM) for feature fusion and extraction, followed by multi-stage support vector regression for later fusion. However, the aforementioned decision-level fusion methods fully consider the differences between modalities and the information within each modality, but they neglect the correlated information between modalities, leading to weak interaction construction among modalities.

Model-level fusion refers to the integration of multiple modalities through corresponding model structures. With the introduction of cross-modal attention mechanisms, attention mechanisms have demonstrated a dominant position in the field of multimodal interaction, making model-level fusion gradually emerge as the mainstream approach in the multimodal domain. Zuo et al. [30] combine video, audio, and text modalities and propose a constraint training strategy based on Central Moment Discrepancy (CMD) distance to extract modality-invariant features. It also introduces invariant features for a missing modality imagination network to predict the invariant characteristics of missing modalities from available ones, thereby reducing the impact of modal gaps. However, this model only considers the invariant features of each individual modality, while neglecting the interactions between them, resulting in weak information interaction capabilities.

In general, model-level fusion is capable of integrating and processing information both within and across models, thereby enabling inter-modal interactions. Additionally, model-level fusion is sufficiently flexible to address the issues present in feature-level fusion and decision-level fusion [8].

In models designed employing the model-level fusion methodology, Zhao et al. [31] propose two modules: modality-specific knowledge-injection and language-guided cross-modal ambiguity learning. The latter uses KL divergence to calculate similarity judgments across different modalities and allocate weights accordingly. Huang et al. [32] introduce two modules: the Causality-Aware Text Debiasing Module and the Counterfactual Cross-modal Attention, employing a Counterfactual attention mechanism. Additionally, the K-means clustering algorithm is used to extract a global dictionary for the text modality, avoiding the situation where the presence of a specific word in the text modality biases the model toward a particular emotion. However, the introduction of the text modality has led to the issue of data redundancy.

Liu et al. [16] propose text information enhancement and multimodal feature fusion, introducing a separate Single-text modality sentiment analysis module for the text modality and presenting a novel cross-modal fusion method to address the existing data redundancy. Goncalves et al. [6] employ a standard multi-head attention mechanism to achieve fusion between modalities and use an auxiliary network to accelerate model convergence. However, these two methods merely facilitate inter-modal information interaction, while overlooking the extraction of modality-invariant features and intra-modal information. This leads to a bias in modality features towards one specific modality, resulting in the loss of information within individual modalities and hindering the completeness of feature learning. Furthermore, Fan et al. [27] rely exclusively on Temporal Pooling in their auxiliary network, which limits the richness of intra-modal feature extraction.

To tackle the aforementioned problems, this paper proposes the CMB module to address the insufficient interaction issue in Zuo et al. [30], and introduces the SAB module to resolve the information omission problem in the study of Fan et al. [27]. Furthermore, we design the AVAM to mitigate issues such as intra-modal data loss and insufficient information extraction observed in Goncalves et al. [6] and Fan et al.'s [27] works. Through these measures, the proposed ATTSF-Net is able to achieve a comprehensive understanding of both intra-modal and inter-modal information, thus facilitating multimodal emotion recognition.

III. Attention-Based Similarity Fusion Network

The research objective of this paper is to perform emotion recognition by leveraging information from audio-visual modalities. Video signal frames are sampled to obtain corresponding video frames, and subsequently, audio-visual features are extracted. The CMB is then employed for modal interaction, while the SAB is utilized for similarity analysis, in conjunction with an AVAM, to achieve final fusion and emotion recognition. The proposed structure is illustrated in Fig. 1. The output of this model is the predicted emotion category. In this study, model-level fusion and a cross-modal attention mechanism are adopted for modal interaction, and the SAB is used to refine the similarity analysis, thereby enhancing the accuracy of emotion recognition. The following steps detail each component of the model, including data preprocessing, feature extraction, CMB, SAB, and AVAM.

A. Preprocessing

For the audio modality, the audio is first sampled at 16,000 Hz in a single channel and then trimmed to a uniform duration of 3.6 seconds. Subsequently, the Mel Frequency Cepstral Coefficients (MFCC) features are extracted using the librosa library.

For the video modality, the video is uniformly sampled to obtain a series of video frames. Each frame is then processed using the MTCNN model for face detection. This model employs three cascaded networks, namely P-Net, R-Net, and O-Net, all based on CNN, to detect faces. It applies different scaling factors to the width and height of the images, progressively reducing them to a fraction of their original size, ensuring that the extracted faces are of a consistent size. A total of 21,600 rectangular face images are cropped from the RAVDESS dataset. These images are subjected to data augmentation techniques, such as random horizontal and vertical flipping, random rotation, and center cropping. The resulting faces are resized to a fixed size of 224x224 pixels for subsequent facial feature extraction.

B. Feature Extraction

For the audio modality, a simple CNN architecture is employed to extract audio features. This architecture is composed of four convolutional layers, with the number of channels set to {64, 128, 256, 128}. The convolutional kernel size, stride and padding are configured as 3, 1, and 0, respectively. Each convolutional layer is followed by a Batch Normalization operation, a ReLU

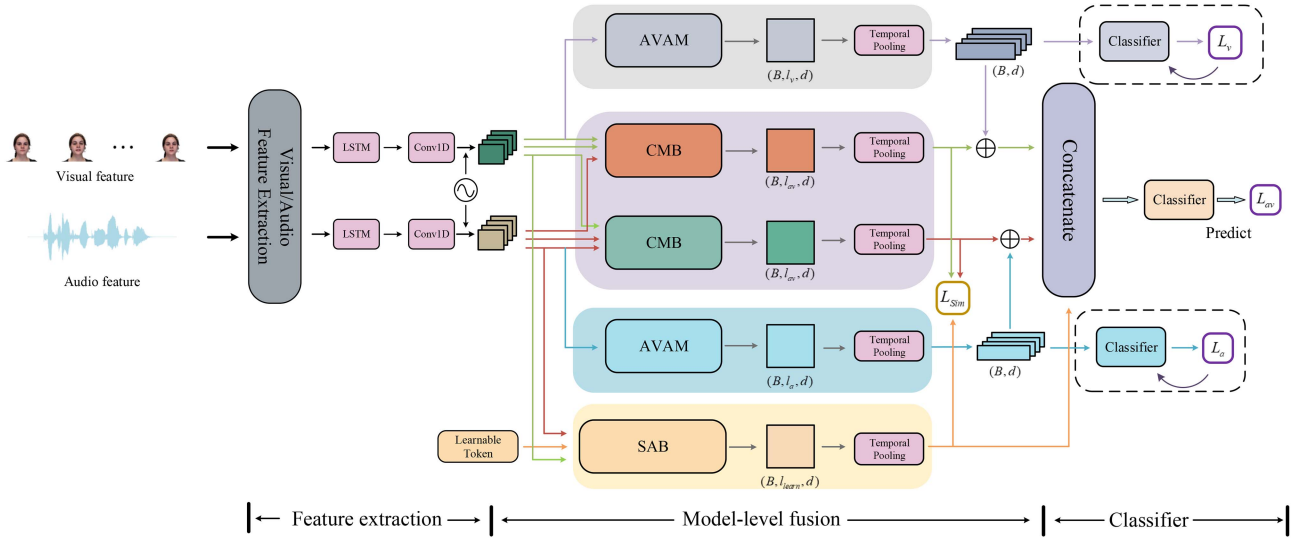

Fig. 1. The schematic diagram of the proposed ATTSF-Net is presented as follows. In this diagram, the lavender block and the pale yellow block represent CMB and SAB, respectively, while both the gray block and the blue block denote AVAM. The two dashed blocks indicate additional loss backpropagation in the AVAM to accelerate model convergence. denotes the positional embedding, and L_{Sim} represents the similarity loss function.

activation function, and a max-pooling layer with a pooling size of (2, 1).

For the video modality, a pre-trained EfficientFace model is leveraged to extract features from video frames [33]. This model, with a reduced number of parameters, ensures the accurate capture of emotional information through a local feature extractor that utilizes deep convolutions and a channel-spatial modulator. Given that most emotions manifest as combinations, mixtures, or compounds of basic emotions, the model incorporates Label Distribution Learning (LDL) as a novel training strategy. The loss function adopts the cross-entropy function to fine-tune the pre-trained EfficientFace model.

Subsequently, LSTM and Conv1D layers are utilized to capture the temporal dependencies and local features of audio features and video features, respectively. The dimensions of both audio and video features are standardized to ensure consistency for subsequent processing. Notably, the concurrent application of Conv1D and LSTM enables the capture of corresponding local temporal features, thereby providing additional informative cues for the subsequent attention mechanism.

C. Cross-Multimodal Block

To fully leverage the information from both audio and video modalities, this paper proposes the CMB for extracting intermodal information. Let a and v represent the corresponding modalities, x_a and x_v represent the inputs from the audio and video modalities, respectively. The dataset is denoted as $D = \{X^i, Y^i\}_i^N$. Where X represents the audiovisual data pairs satisfying $X^i = \{x_a, x_v\}_i$. N is the number of samples. Y denotes the corresponding emotional labels with o categories, d be the feature dimension, and l be the sequence length of the respective modalities. The objective of the proposed model is to construct a function that satisfies $f(x_a, x_v) \rightarrow Y$, where the right arrow denotes the mapping relationship of the model. The architecture employs a cross-modal attention mechanism as its backbone. It consists of two branches: an audio branch and a video branch. Each branch utilizes the cross-modal attention mechanism to compute hidden representations across the two modalities, which can be expressed as:

$$x_{a\rightarrow v} = f_{CMB}^{a\rightarrow v}\left(x_a, x_v, \theta_{CMB}^{a\rightarrow v}\right) \quad x_{a\rightarrow v} \in \mathcal{R}^{d\times l} \tag{1}$$

$$x_{v\rightarrow a} = f_{CMB}^{v\rightarrow a}\left(x_v, x_a, \theta_{CMB}^{v\rightarrow a}\right) \quad x_{v\rightarrow a} \in \mathcal{R}^{d\times l} \tag{2}$$

Here, x_a represents the audio modality, x_v represents the video modality, f_{CMB} denotes the CMB, and θ represents the parameters contained within the block. $a \rightarrow v$ and $v \rightarrow a$ represent two scenarios where the audio modality and the video modality dominate, respectively. The CMB facilitates interaction between modalities through an improved cross-modal attention mechanism. The specific structure is illustrated in Fig. 2, where one modality is used as the Query in the cross-modal attention, interacting with the Key and Value of the other modality. Taking the scenario where the video modality dominates as an example, the specific formula can be expressed as:

$$Q_v = W_q^v x_v \tag{3}$$

$$K_v = W_k^v x_v \tag{4}$$

$$V_a = W_v^a x_a \tag{5}$$

In this formula, W_q, W_k and W_v represent learnable parameters, x_a and x_v denote the features extracted from the audio and video modalities, respectively. The attention mechanism [11] is then used for computation, with the formula given by:

$$att_{v\rightarrow a} = Softmax\left(\frac{Q_v K_a^T}{\sqrt{d}}\right) V_a \tag{6}$$

$$\bar{x}_{v\rightarrow a} = LN(x_v + att_{v\rightarrow a}) \tag{7}$$

$$\tilde{x}_{v\rightarrow a} = LN(Multi - Head(\bar{x}_{v\rightarrow a}) + \bar{x}_{v\rightarrow a}) \tag{8}$$

$$x_{v\rightarrow a} = LN(FFN(\tilde{x}_{v\rightarrow a}) + \tilde{x}_{v\rightarrow a}) \tag{9}$$

Here, Q_v, K_a, V_a represent the Query, Key, and Value from the two modalities, respectively. LN denotes the LayerNorm layer,

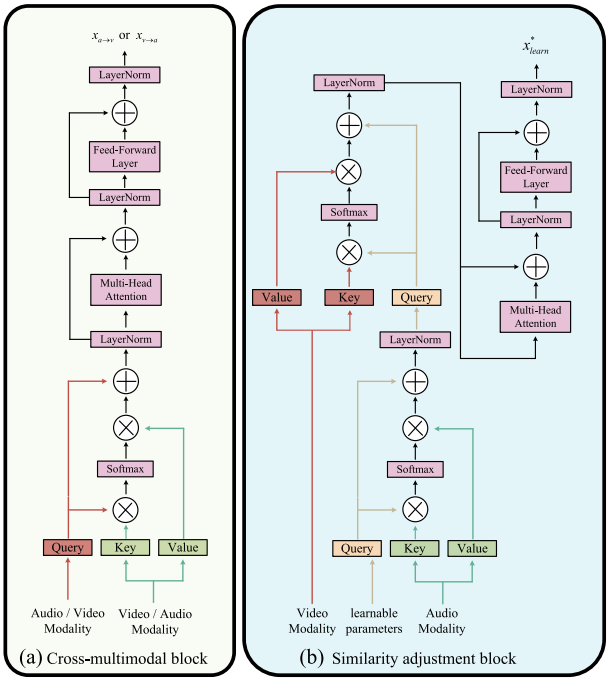

Fig. 2. The structural diagrams for the CMB and the SAB.

$Multi-Head$ represents the standard multi-head attention mechanism, FFN stands for a feed-forward network (FFN) used to achieve nonlinearity, and the scaling factor d in the Softmax function is employed to mitigate potential gradient vanishing issues caused by the Softmax operation.

D. Similarity Adjustment Block

To enable the model to learn modal-invariant features and further deepen the emotional information across modalities, this paper proposes the SAB. In most Transformer-based classification tasks, such as the Vision Transformer (VIT) model [34] for images and the Audio Spectrogram Transformer (AST) model [35] for audio, a learnable CLS token is utilized and embedded into the feature sequence for attention computation, followed by category classification based on this CLS token. Inspired by this idea, the SAB introduces a learnable token to facilitate multi-modal information learning and incorporates a similarity function to ensure the model captures the correct emotional information. The detailed structure of this block is illustrated in Fig. 2. This structure can be represented as:

$$x^*_{learn} = f_{SAB}(x_{learn}, x_a, x_v, \theta_{SAB}) \tag{10}$$

In the formula, f_{SAB} denotes the SAB, x_a and x_v represent the corresponding audio and video modalities, respectively, θ represents the parameters contained within the structure, and x^*_{learn} denotes the learned emotional information. This structure initially utilizes a learnable token as the Query in the attention mechanism. Subsequently, it performs attention computations with the Key and Value from both the audio modality and the video modality to learn emotional information from each. This is followed by a multi-head attention block, and then a FFN is employed to achieve nonlinearity. This structure can be described as:

$$\bar{x}_{learn} = LN(att_{learn\to a}(att_{learn\to v})) \tag{11}$$

$$\tilde{x}_{learn} = LN(Multi-Head(\bar{x}_{learn}) + \bar{x}_{learn}) \tag{12}$$

$$x^*_{learn} = LN(FFN(\tilde{x}_{learn}) + \tilde{x}_{learn}) \tag{13}$$

In the formula, $att_{learn\to a}$ and $att_{learn\to v}$ represent the emotional information learned by the learnable token from the audio and video modalities, respectively. LN denotes the Layer Normalization layer, $Multi-Head$ represents the standard multi-head attention mechanism, and FFN stands for a FFN. When the dominant modality and the secondary modality are consistent, the fused modality can reinforce the dominant one. However, when there is a discrepancy between the dominant and secondary modalities, it may introduce additional noise. To alleviate this issue, we take into account the posterior distributions obtained from the CMB and SAB.

After being processed by the CMB, the features first undergo temporal pooling, resulting in features denoted as $x_k \in \mathcal{R}^d$. At this stage, the video and audio modalities are expected to exhibit identical data distributions. Notably, since these features are represented as vectors, we can compute their distribution. Given a sufficiently large dataset, it is assumed that both modalities follow a normal distribution.

$$x_k \sim \mathcal{N}\left(x_k; \mu_k, \sigma_k^2\right) \quad k \in \{a \to v, v \to a\} \tag{14}$$

To quantify the similarity between the posterior distributions of the outputs obtained from CMB and SAB, we utilize the KL divergence as a metric to measure the disparity between the outputs of CMB and SAB. Let P_m represent the distribution of $x_{m\to m'}$, where m and m' denote different modalities. Thus, the similarity loss function can be calculated as:

$$L_{a-l} = D_{kl}(P_{learn} \parallel P_a) + D_{kl}(P_a \parallel P_{learn}) \tag{15}$$

$$L_{v-l} = D_{kl}(P_{learn} \parallel P_v) + D_{kl}(P_v \parallel P_{learn}) \tag{16}$$

In the formula, $D_{kl}(.)$ represents the KL divergence, and a and v represent the audio and video modalities, respectively. Let L_{a-l} and L_{v-l} denote the similarity metrics between the multimodal sentiment features extracted from SAB and the audio or video features obtained from the CMB, respectively. A smaller value of these metrics indicates a higher degree of similarity between the two sets of features. To ensure that the features extracted by the CMB and the SAB are highly similar, we introduce a similarity loss function, which can be mathematically expressed as follows:

$$L_{sim} = \frac{1}{2}(L_{a-l} + L_{v-l}) \tag{17}$$

E. Audio-Visual Auxiliary Modules

To fully capture emotional information while preventing intra-modal information loss during inter-modal feature extraction, we propose the AVAM. The AVAM provides an additional semantically meaningful representation layer for emotion recognition, thereby enhancing deep neural network performance.

In the model proposed in this paper, each AVAM learns emotional information from a single modality and is fused with

the main network through a fusion strategy designed in Section F of this chapter. The detailed process is illustrated in Fig. 1. Since the input of AVAM is solely from a single modality and only exchanges data with the main network during the fusion process, they are modality-independent. The AVAM ensures that the model still contains layers with complete information even when one modality is missing.

The proposed AVAM adopts a standard Transformer Encoder structure, comprising a multi-head attention mechanism and a FFN. The multi-head attention mechanism ensures effective single-modality learning, while the FFN performs nonlinear transformation. Finally, average pooling integrates the auxiliary network's features, thus preserving intra-modal information during training and inference. Additionally, two cross-entropy loss functions optimize the AVAM, accelerating training and guaranteeing proper auxiliary network learning.

The complete framework of the proposed model consists of three modules: CMB, SAB, and AVAM, each equipped with its own cross-entropy loss L_m. where $m \in \{a, v, av\}$. Here, a represents the audio auxiliary module, v represents the video auxiliary module, and av represents the backbone network. This model explores a mechanism for learning shared information through losses in multiple network streams, thereby learning multimodal representations. Taking into account the L_{sim} in the SAB, the total loss function L_{total} is obtained by combining the corresponding loss functions as follows:

$$L_{total} = L_{av} + \lambda_a L_a + \lambda_v L_v + \lambda_{sim} L_{sim} \tag{18}$$

In the formula, L_{av}, L_a, L_v, and L_{sim} represent the corresponding loss functions from the CMB, auxiliary networks, and SAB, respectively. λ_a, λ_v, and λ_{sim} denote the respective hyperparameter weights.

F. Fusion Strategy

In this subsection, we explore different fusion strategies for the proposed ATTSF model. Inspired by references [8], [27], and [9], we design three methods: ATTSF-Net, ATTSF-L2, and ATTSF-Gat. These approaches are based on element-wise summation, L2 norm, and a learnable parameter layer for feature fusion, respectively.

ATTSF-Net employs element-wise summation for feature fusion. Taking the audio modality as an example, given the hidden features obtained from the main network and the auxiliary network, the final feature can be expressed as:

$$x^*_{v\to a} = x_{v\to a} + x^*_v \tag{19}$$

where $x_{v\to a}$ and x^*_v denote the hidden features obtained from the main network and the auxiliary network, respectively, and $x^*_{v\to a}$ represents the final feature obtained for this branch.

ATTSF-L2 utilizes a fusion method based on the $L2$ norm. After obtaining the hidden features from the main network and the auxiliary network, the corresponding weights are calculated using the $L2$ norm. The calculation formula is as follows:

$$W_{av} = \frac{\| H_{av} \|}{\| H_{av} \| + \| H_m \|} \tag{20}$$

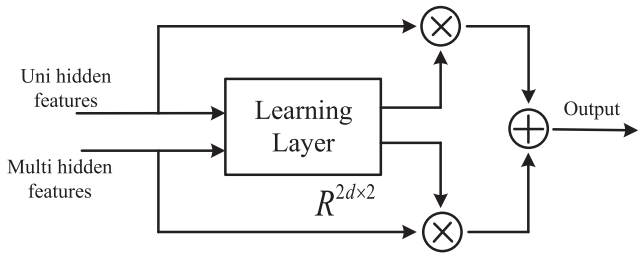

Fig. 3. The diagram of the Learning Layer fusion method.

$$W_m = \frac{\| H_m \|}{\| H_{av} \| + \| H_m \|} \tag{21}$$

Here W_{av} represents the weight of the main network, and W_m denotes the weight of the auxiliary module, and H_{av} and H_m denote the features processed by the multimodal branch and unimodal branch, respectively. where $m \in \{a, v\}$, $||\cdot||$ represents the L2 norm.

ATTSF-Gat trains the weights of the main network and the auxiliary network using a simple linear layer, named the Learning Layer. The specific implementation of this layer is illustrated in Fig. 3. The input dimension is $2d$ and the output dimension is 2. This layer combines the hidden features from the main network and the auxiliary network as input. The two output dimensions represent the weight parameters for the unimodal hidden features and the multimodal hidden features, respectively.

IV. Experiment Settings

This section will describe the dataset used and the details of model training.

A. Datasets

RAVDESS: The Ryerson Audio-Visual Database of Emotional Speech and Song comprises 24 professional actors (12 males and 12 females) speaking two lexically-matched statements in a neutral North American accent. The dataset includes both speech and song components. The speech component encompasses eight emotional categories: 'calm', 'happy', 'sad', 'angry', 'surprised', 'disgusted', 'fearful', and 'neutral', each expressed at two levels of emotional intensity (normal and strong), along with a neutral emotion. The singing component includes emotional categories such as 'calm', 'happy', 'sad', and 'angry'. In this experiment, we utilize only the audio-visual data from the speech component of RAVDESS, totaling 1440 samples. The data distribution is shown in Table I. Consistent with the methodology outlined in [27], we split the data into training, validation, and test sets in a 6:2:2 ratio and employed 5-fold cross-validation.

IEMOCAP: The Interactive Emotional Dyadic Motion Capture dataset is collected by the Speech Analysis and Interpretation Laboratory at the University of Southern California and is widely used for multimodal emotion recognition. The dataset captures the performances of ten actors in dyadic conversations, where they act out selected emotional scripts and improvised scenarios intended to elicit specific emotions. The database

TABLE I
DISTRIBUTION CHARTS OF EMOTION CATEGORIES IN THE RAVDESS AND IEMOCAP DATASETS

Emotion Categories	Neutral	Happy	Sad	Angry	Calm	Fearful	Disgust	Surprised
RAVDESS	192	384	384	384	384	384	384	384
IEMOCAP	1708	1636	1084	1103	-	-	-	-

TABLE II
PARTIAL PARAMETER TABLE FOR MODELS USING RAVDESS AND IEMOCAP DATASETS

-	RAVDESS	IEMOCAP
Feature Dimension	128	128
Sequence Length	128	64
Number of CMB Layers	6	5
Number of SAB Layers	6	5
Number of Auxiliary Network Layers	2	1

TABLE III
COMPARISON TABLE OF ACCURACY ON RAVDESS DATASET

Model	Years	UA.
Ghaleb et al. [36]	2019	67.70
MMTM [37]	2020	73.12
Ghaleb et al. [38]	2020	76.30
TA-AVN [5]	2021	78.70
Luna-Jimene et al. [39]	2021	80.08
ERANNs [40]	2022	74.80
Guo et al. [41]	2022	78.45
Chumachenko et al. [42]	2022	81.58
V8+A4 [43]	2022	86.00
FT-Wav2Vec+bi-LSTM+MLP [44]	2022	86.70
AttN-NET (end-to-end) [27]	2024	82.62
AttN-NET (Non-end-to-end) [27]	2024	88.00
Human [22]	2018	80.00
ATTSF-Gat	2025	85.50
ATTSF-L2	2025	87.87
ATTSF-Net	2025	**88.67**

features five dyadic interactive dialogues, each lasting approximately five minutes and segmented at the sentence level. Each utterance is annotated with a single emotional category label. In this experiment, we conduct a four-class classification task using four emotional categories: 'Neutral', 'Happy', 'Sad', and 'Angry'. The dataset comprises a total of 5,531 data samples. The data distribution is shown in Table I. Consistent with the approach in [15], we use the first four sessions as the training set and the last session as the test set.

B. Details of Model Training

The model trains using NVIDIA RTX 2080Ti GPUs (11 GB) with the PyTorch framework. The Adam optimizer is employed, and the mini-batch sizes are set to 8 and 32 for the RAVDESS and IEMOCAP datasets, respectively. The learning rates are set to 0.0004 and 0.001, respectively, and the model is trained for 200 epochs in both cases. Starting from the 30th epoch, the learning rate is decayed by 0.9 every 20 epochs. The weight decay is set to 0.00001, and we employ the Early Stopping method. Experimental results are reported on the test set.

Previous research has shown that using all frames of video data often leads to an increased computational load without necessarily increasing the emotional information contained. Therefore, in this experiment, 15 frames are uniformly selected from each video as the video modality. For the MFCC in the audio modality, a Hamming window with a window length of 50 ms and a frame shift of 30 ms is used, and the first 12 coefficients are selected as input. Regarding model parameters, on the RAVDESS dataset, the number of neurons in the FFN layers of CMB, SAB, and AVAM is set to 256. On the IEMOCAP dataset, the number of neurons in the FFN layers of CMB and SAB is 256, while the number of neurons in the FFN layer of the AVAM is 128. Other model parameters are shown in Table II. To prevent overfitting, Dropout layers are added between each layer, and dropout is also applied to the data in the main path during residual connections, with a dropout rate of 0.4 for both. In the loss function, the coefficients λ_a, λ_v, and λ_{sim} are set to 0.5, 0.5, and 0.3, respectively.

Notably, the RAVDESS dataset comprises only audio and video modalities, whereas the IEMOCAP dataset includes audio, video, and text modalities. To maintain consistency in our model, we utilize only the audio and video modalities. However, to ensure a fair comparison, we conduct additional experiments on the IEMOCAP dataset by incorporating the text modality. We employ a pre-trained BERT model to extract text features, following the same data processing methodology as in the AVAM module. The text modality is integrated only at the concatenate stage of our model, with no backbone network computations involved in other parts.

V. EXPERIMENT RESULTS AND DISCUSSION

A. Comparison With Multimodal Baselines

In this section, we compare our proposed model with multimodal baseline models. The models are trained using the standards and optimization methods discussed in Section IV. First, Tables III and IV report the accuracy of different models on the RAVDESS and IEMOCAP datasets, respectively. The results indicate that our proposed ATTSF-Net achieves superior performance. For the RAVDESS dataset, our model attains the highest accuracy, outperforming the AttN-NET (Non-end-to-end) model by 0.67 percentage points and achieving a maximum accuracy of 88.67%. And our proposed model achieves performance surpassing that of humans, further proving its superiority. After utilizing all three modalities, we achieve the highest UA and WA accuracies of 81.93% and 79.77%, respectively. Our model achieves the best results in terms of UA accuracy, while the ATIA model performs best in terms of WA accuracy. Overall, the ATTSF model, integrating CMB and SAB modules, exhibits superior performance on both RAVDESS and IEMOCAP datasets.

TABLE IV
COMPARISON TABLE OF ACCURACY ON IEMOCAP DATASET

Model	Years	Modality	WA.	UA.
Li et al. [45]	2018	A+T	72.1	71.9
CRNN [46]	2019	A	68.78	64.16
CNN-TF-GAP [46]	2019	A	73.33	64.80
Xu et al. [47]	2020	A+T	72.5	70,90
Liang et al. [48]	2020	A+T+V	75,6	74.5
Kumar et al. [49]	2021	A+T	71.70	75.00
Transfer Learning [50]	2022	A+T	75.76	76.07
Transfer Learning [50]	2022	A	65.40	65.97
Transfer Learning [50]	2022	T	70.24	70.33
ATIA [15]	2022	A+T	**82.40**	80.60
IF-MMIN [51]	2023	A+V	65.33	66.52
IF-MMIN [51]	2023	A+T	73.05	75.44
Khan et al. [28]	2024	A+V	80.56	79.69
MemoCMT [29]	2025	A+T	81.85	81.33
ATTSF-Gat	2025	A+V	76.01	79.21
ATTSF-L2	2025	A+V	77.09	79.95
ATTSF-Net	2025	A+V	76.61	79.90
ATTSF-Gat	2025	A+V+T	78.92	81.23
ATTSF-L2	2025	A+V+T	79.04	81.39
ATTSF-Net	2025	A+V+T	79.77	**81.93**

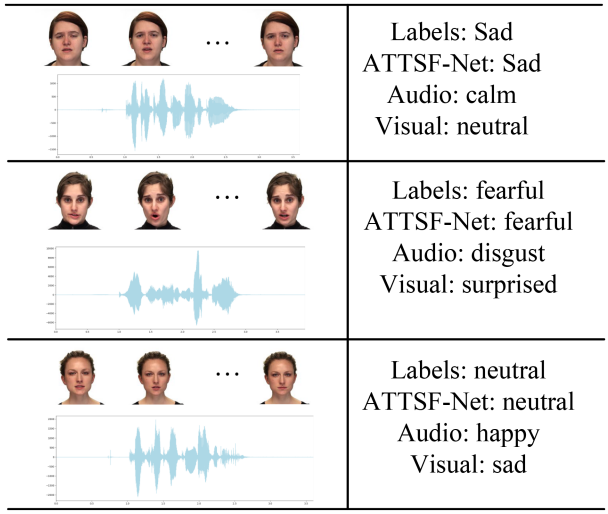

Fig. 4. Examples of Emotion Recognition Results Using Multimodal Versus Unimodal Data.

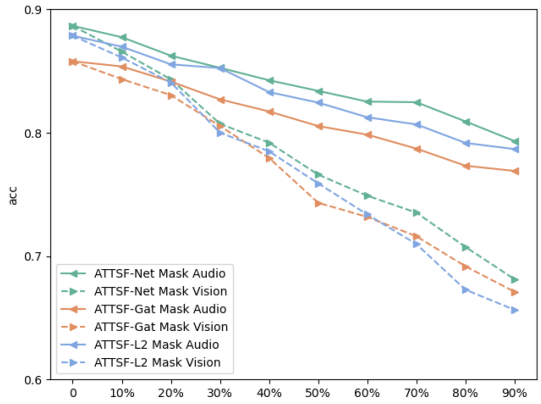

Fig. 5. Masking Modality Data Experiments of ATTSF-Gat, ATTSF-L2, and ATTSF-Net Models in the RAVDESS Dataset.

B. Evaluation of the ATTSF

1) Multimodal Effectiveness Experiment: In this section, we report the comparative results of unimodal and multimodal approaches on the RAVDESS and IEMOCAP datasets, as detailed in Table VI. Specifically, on the RAVDESS dataset, the multimodal classification results outperform the unimodal audio and video modalities by 10 and 12 percentage points, respectively. On the IEMOCAP dataset, for UA accuracy, the multimodal approach surpasses the unimodal ones by 4.53 and 10.39 percentage points, respectively, demonstrating the superiority of multimodal signals over unimodal data in emotion recognition. Taking the RAVDESS dataset as an example, in the 5th fold, the labels for samples with IDs 3, 113, and 128 are 'sad', 'fearful', and 'neutral', respectively. However, the audio data misclassifies these emotions as 'calm', 'disgust', and 'happy', while the video data misclassifies them as 'neutral', 'surprised', and 'sad'. Only the multimodal approach correctly classifies the emotions, as illustrated in Fig. 4. These discussions highlight that multimodal data outperforms unimodal data in emotion recognition by effectively fusing emotional information from individual modalities. Notably, even when using only video modality data on the RAVDESS dataset, our approach still outperforms approximately half of the baseline models. Similarly, when using only audio modality data, we also outperform about half of the baselines, indicating the robust stability of our proposed ATTSF-Net. This aspect is further confirmed by the robustness experiments presented in the next section.

2) Model Robustness Experiment: In this subsection, we evaluate the robustness performance of the ATTSF-Net, ATTSF-L2, and ATTSF-Gat models on the RAVDESS dataset, respectively. Specifically, we randomly overlay a portion of the audio and video data with random values to simulate scenarios with random noise. The coverage rate ranges from 10% to 90%, increasing in increments of 10%. For instance, a 30% mask implies that 30% of the audio or video data is randomly replaced with random values. This evaluation aims to analyze the robustness of the models when only partial information from one modality is available.

The experimental results are shown in Fig. 5 and Table V. In the figure, the green, blue, and orange lines represent the ATTSF-Net, ATTSF-L2, and ATTSF-Gat models, respectively. Solid lines indicate audio data masking, while dashed lines indicate video data masking. The horizontal axis represents the masking ratio, and the vertical axis represents the model accuracy. As observed from the figure, the accuracy of all three models decreases as the masking ratio increases. However, ATTSF-Net consistently outperforms the other two models in terms of accuracy. Notably, when the masking ratio of audio and video data reaches 90%, the accuracy of all three models tends towards their respective unimodal accuracies. When the data masking ratio is between 20% and 70%, the accuracy of the ATTSF-Net model decreases relatively slowly. Specifically, in the range of 60% to 70%, the model's accuracy drops from 82.53% to 82.48%, a mere decrease of 0.05 percentage points. The accuracy of ATTSF-Net with less than 20% of video data masked remains higher than that of ATTSF-Gat. This indicates that the ATTSF-Net, which employs element-wise summation

TABLE V
ACCURACY TABLE FOR MODALITY MASKING EXPERIMENTS IN RAVDESS DATASETS

Masking Rate	10%	20%	30%	40%	50%	60%	70%	80%	90%
ATTSF-Net(audio)	0.8775	0.8625	0.8527	0.8426	0.8340	0.8253	0.8248	0.8092	0.7933
ATTSF-Net(visual)	0.8658	0.8433	0.8073	0.7920	0.7661	0.7490	0.7353	0.7075	0.6813
ATTSF-L2(audio)	0.8698	0.8555	0.8525	0.8330	0.8245	0.8126	0.8068	0.7918	0.7868
ATTSF-L2(visual)	0.8610	0.8408	0.8001	0.7850	0.7590	0.7338	0.7098	0.6728	0.6565
ATTSF-Gat(audio)	0.8538	0.8415	0.8271	0.8173	0.8055	0.7985	0.7872	0.7733	0.7691
ATTSF-Gat(visual)	0.8435	0.8305	0.8055	0.7796	0.7433	0.7318	0.7161	0.6918	0.6711

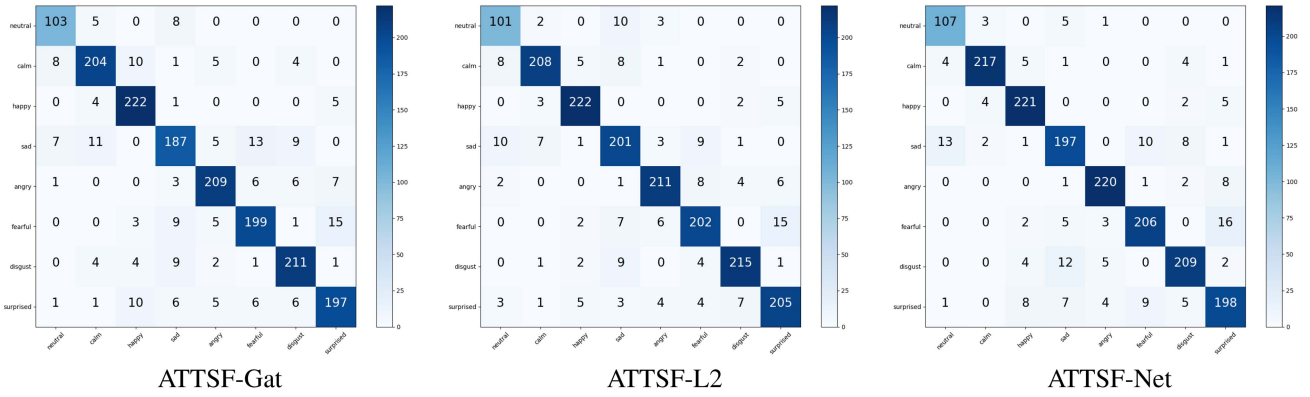

Fig. 6. The confusion matrix of the RAVDESS dataset.

TABLE VI
COMPARISON TABLE OF UNIMODAL VERSUS MULTIMODAL

Modality	UA. (RAVDESS)	WA. (IEMOCAP)	UA. (IEMOCAP)
ATTSF-Net	**88.67**	**79.77**	**81.93**
ATTSF-Net (Audio)	66.43	74.37	77.40
ATTSF-Net (Visual)	78.50	66.62	71.54

TABLE VII
TABLE OF ABLATION STUDY ACCURACY, WHERE "W/O" REFERS TO "WITHOUT" THE RELEVANT COMPONENT

Model	UA. (RAVDESS)	WA. (IEMOCAP)	UA. (IEMOCAP)
W/o $Loss_{sim}$	87.43	78.02	80.38
W/o Audio auxiliary	85.65	79.28	81.50
W/o Visual auxiliary	88.46	79.50	81.87
W/o SAB module	86.86	77.88	80.27
ATTSF-Net	**88.67**	**79.77**	**81.93**

for fusion, exhibits greater robustness compared to the ATTSF-Gat that utilizes Learning Layer for fusion.

Additionally, the figure shows that ATTSF-Gat exhibits greater robustness than ATTSF-L2 when more than 60% of the video data is masked. The analysis demonstrates that our proposed method is robust to missing modal data, and its performance consistently improves as more audio or video features become available. When the amount of missing data in audio or video features is between 10% and 30%, our model's performance approaches that of a complete audiovisual scenario.

3) Visual Analysis: In this section, we employ a series of visualizations to analyze the proposed models. Fig. 6 presents the confusion matrices of the proposed ATTSF-Gat, ATTSF-L2, and ATTSF-Net on the RAVDESS dataset. As observed from the figure, the classification results of ATTSF-Net are more distinct compared to the other two models. Specifically, ATTSF-Net exhibits lower classification accuracy in the three emotion categories of 'surprised', 'disgust', and 'happy' than the other two models, but achieves higher classification accuracy in the remaining five emotion categories. Furthermore, ATTSF-Net basically achieves classification across the RAVDESS dataset.

Figs. 7 and 8 display the t-SNE visualizations of the proposed ATTSF-Net on the RAVDESS and IEMOCAP datasets, respectively. From the figures, it can be seen that for the RAVDESS dataset, the output distributions of the three models are generally consistent. However, ATTSF-Gat demonstrates a poorer distribution of the 'fearful' category, with some confusion occurring between the 'fearful' category and the 'sad', 'disgust' categories. ATTSF-L2 also results in some data adhesion, particularly among the 'sad', 'surprised', and 'angry' categories, while it performs better in classifying 'happy' and 'calm'. ATTSF-Net achieves the best results, with only slight adhesion between 'angry' and 'surprised'. For the IEMOCAP dataset, the ATTSF-Gat and ATTSF-L2 models exhibit poorer classification performance for the 'Neural' category, as evidenced by significant confusion between the 'Neural' category and other categories in the figures. In contrast, the ATTSF-Net model is able to identify the centers of data point clusters, achieving clear clustering of different categories. In summary, the proposed ATTSF-Net significantly distinguishes between different categories.

C. Ablation Study

In this section, we conduct a series of ablation experiments on the proposed ATTSF-Net, where we sequentially remove the $Loss_{sim}$, audio auxiliary network, video auxiliary network, and SAB. The experimental results are presented in Table VII. As shown in Table VII, within the RAVDESS dataset, the audio

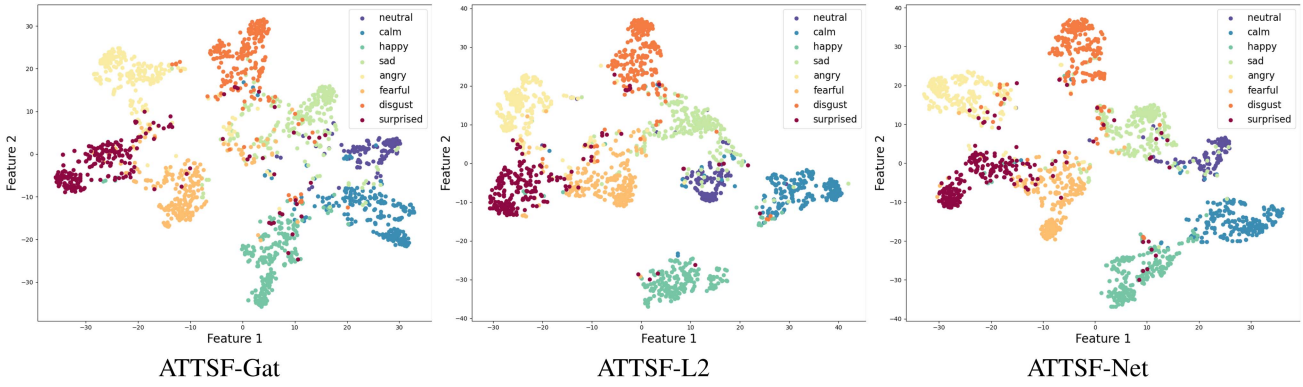

Fig. 7. t-SNE visualization of ATTSF on the RAVDESS dataset.

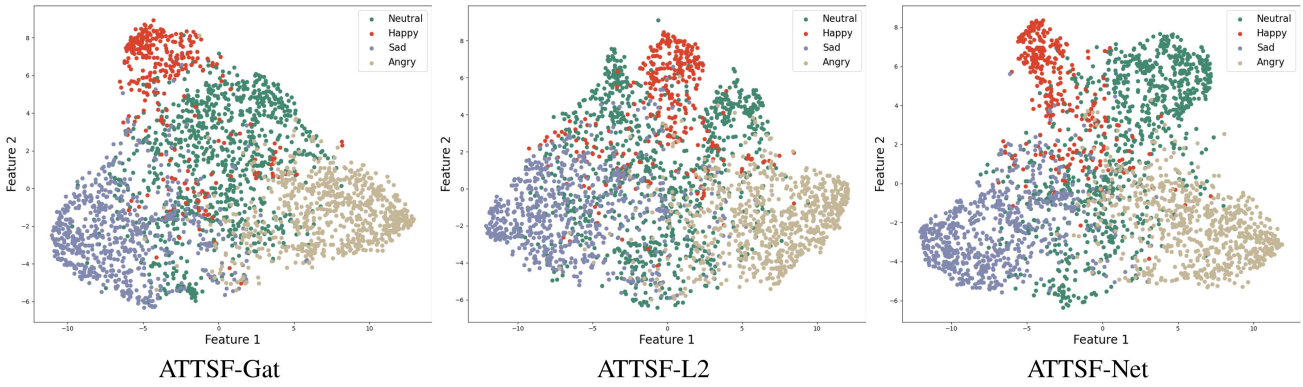

Fig. 8. t-SNE visualization of ATTSF on the IEMOCAP dataset.

auxiliary network has the most significant impact, enhancing the model's performance by 3.02 percentage points. The SAB follows closely, boosting the model's accuracy by 1.81 percentage points. Moreover, the incorporation of this module enables ATTSF-Net to outperform the ATTN model [27] in non-end-to-end scenarios. Notably, the video auxiliary network only contributes to a 0.21 percentage point improvement, which is relatively minor compared to the audio auxiliary network. We hypothesize that this is because the audio signals in the dataset contain richer information, and the audio auxiliary network effectively integrates intra-modal and inter-modal emotional information from the audio stream, further enhancing the model's accuracy.

$Loss_{sim}$ improves the model by approximately one percentage point. In the IEMOCAP dataset, the SAB and $Loss_{sim}$ are the most beneficial to the model, each contributing to an improvement of about 1.5 percentage points. Furthermore, the incorporation of the SAB module enables the model to outperform the approach in reference [15] in terms of UA accuracy, indicating that the designed SAB and $Loss_{sim}$ allow the model to fully consider both intra-modal and inter-modal emotional information. Even when the SAB is removed, the proposed ATTSF-Net still outperforms the end-to-end scenarios presented in references [27], [43], and [44] within the RAVDESS dataset. This suggests that the CMB included in ATTSF-Net is sufficiently effective for audio-video emotion recognition.

TABLE VIII
EXPERIMENTS ON MODULES REPLACEMENT ON THE RAVDESS DATASET, WHERE "→" DENOTES THE REPLACEMENT OPERATION

Model Combination	UA.
CMB → Similarity-aware Enhancement Block [27]	86.67
AVAM → Temporal Pooling [27]	85.81
LSTM → Linear	87.25
ATTSF-Net	**88.67**

Both the similarity-aware enhancement block and temporal pooling components are sourced from the attn model [27], and "Linear" represents a Linear layer.

The experiments demonstrate that the model's performance deteriorates when these designs are excluded, confirming that our approach is effective as expected.

We conduct experiments to replace modules within our model to validate their effectiveness, with the results presented in Table VIII. Initially, inspired by the ATTN model [27], we replace the CMB and AVAM modules in our model with the Similarity-aware Enhancement Block and Temporal Pooling mechanisms from ATTN, respectively. However, this modification leads to a decline in accuracy. Additionally, we substitute the LSTM layer with a simple linear layer, which results in an approximately 1% drop in accuracy. This indicates that the LSTM layer, while adjusting features to a fixed dimension, further integrates local temporal features, thereby enhancing model performance.

TABLE IX
THE EXPERIMENTAL RESULTS OF ATTSF AND PREVIOUS STUDY [17] ON THE ADDITIONAL CMU-MOSI DATASETS, WHERE ACC-2 AND F1 DENOTES THE POSITIVE/NEGATIVE CLASSIFICATION

Model	Acc-2	Acc-7	F1
RAVEN*	78.00	33.20	76.60
MuLT*	81.50	40.00	80.60
BIMHA*	80.20	35.90	80.20
DSIN*	83.10	42.90	83.20
ATTSF-Gat	83.82	40.00	83.66
ATTSF-L2	84.26	43.58	84.15
ATTSF-Net	**84.45**	**46.94**	**84.39**

(* denotes the results collected from the previous study)

D. Case Study

Furthermore, to validate the generalizability of our proposed method, we conduct additional analyses on the CMU-MOSI sentiment analysis dataset and compare our results with those of previous models [17]. The CMU-MOSI dataset is designed for sentiment regression analysis, where each sentiment sample is mapped to a numerical range within [-3,3], with negative values indicating negative sentiment and positive values indicating positive sentiment. We employ the same evaluation metrics as those used in the DSIN model, including binary accuracy (Acc-2), seven-class accuracy (Acc-7), and weighted F1 score (F1). We report the performance on the test set to ensure a fair comparison. The processing approach for the textual modality remained consistent with that used on the IEMOCAP dataset, and the L1 loss function was utilized.

Table IX presents a comparison of the performance metrics between our proposed model and the previous model [17]. It is evident that our proposed ATTSF-Net achieves the optimal performance, surpassing the DSIN model by 1.3 percentage points in Acc-2 and 1.1 percentage points in F1, respectively. ATTSF-L2 achieves the second-best performance, while ATTSF-Gat slightly outperforms the previous model.

When compared with all models in previous studies, these results indicate that our proposed ATTSF-Net achieves favorable outcomes on both classification-based and regression-based datasets. The three distinct fusion mechanisms demonstrate greater effectiveness and robustness, thereby validating the generalizability of our proposed method.

VI. CONCLUSION

This paper introduces a novel multimodal audiovisual fusion model named ATTSF-Net. which comprises the proposed CMB, SAB, and AVAM. The objective is to achieve multimodal emotion recognition. CMB employs cross-modal attention mechanisms and model-level fusion to facilitate interactions between modalities. SAB enables the model to learn modality-invariant features, while AVAM utilizes auxiliary networks to provide additional intra-modal emotional information. To ensure the consistency of the learned audio-visual emotional information, a similarity loss function based on KL divergence is designed and used for model training. Furthermore, two variants of the model, namely ATTSF-L2 and ATTSF-Gat, are proposed using L2 norm fusion and Learning Layer fusion methods, respectively.

These models are validated on the RAVDESS and IEMOCAP datasets. Experimental results demonstrate that the proposed ATTSF-Net achieves the highest accuracy of 88.67% on the RAVDESS dataset and 81.93% UA and 79.77% WA on the IEMOCAP dataset, showcasing the superiority of the proposed model. Additionally, a series of robustness and visualization experiments, including confusion matrix visualization and t-SNE visualization, are conducted. The results indicate that ATTSF-Net can adequately capture intra-modal and inter-modal emotional information, ensuring comprehensive information collection while reducing data redundancy. Ablation experiments further confirm the effectiveness of the proposed SAB module and $Loss_{sim}$, enhancing the accuracy of multimodal emotion recognition.

Although ATTSF-Net can capture inter-modal relationships, it still has some limitations. First, ATTSF-Net relies solely on the interaction between audio and video data. This may lead to inadequate handling of emotions when some modalities are missing, resulting in incomplete information acquisition. Second, ATTSF-Net achieves interaction based on a cross-modal attention mechanism, which demands substantial computational resources, posing deployment challenges on small-scale devices. In future research, we aim to further improve ATTSF-Net by increasing the number and quality of the fused modalities. Additionally, we will strive to reduce the model's complexity through the integration of advanced algorithms, endeavoring to make the model more intelligent in the field of emotion recognition.

REFERENCES

[1] S. Anand, N. K. Devulapally, S. D. Bhattacharjee, and J. Yuan, "Multi-label emotion analysis in conversation via multimodal knowledge distillation," in *Proc. 31st ACM Int. Conf. Multimedia*, New York, NY, USA, 2023, pp. 6090–6100.
[2] J.-H. Hsu and C.-H. Wu, "Applying segment-level attention on bi-modal transformer encoder for audio-visual emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 14, no. 4, pp. 3231–3243, Oct.-Dec. 2023.
[3] T. Song, W. Zheng, P. Song, and Z. Cui, "EEG emotion recognition using dynamical graph convolutional neural networks," *IEEE Trans. Affect. Comput.*, vol. 11, no. 3, pp. 532–541, Jul.-Sep. 2020.
[4] M. Tran and M. Soleymani, "A pre-trained audio-visual transformer for emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 4698–4702.
[5] A. Radoi, A. Birhala, N.-C. Ristea, and L.-C. Dutu, "An end-to-end emotion recognition framework based on temporal aggregation of multimodal information," *IEEE Access*, vol. 9, pp. 135559–135570, 2021.
[6] L. Goncalves and C. Busso, "Robust audiovisual emotion recognition: Aligning modalities, capturing temporal information, and handling missing features," *IEEE Trans. Affect. Comput.*, vol. 13, no. 4, pp. 2156–2170, Oct.-Dec. 2022.
[7] Z. Zhao, Y. Wang, G. Shen, Y. Xu, and J. Zhang, "TDFNet: Transformer-based deep-scale fusion network for multimodal emotion recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 3771–3782, 2023.
[8] H. Zhou, J. Du, Y. Zhang, Q. Wang, Q.-F. Liu, and C.-H. Lee, "Information fusion in attention networks using adaptive and multi-level factorized bilinear pooling for audio-visual emotion recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 2617–2629, 2021.
[9] R. G. Praveen and J. Alam, "Incongruity-aware cross-modal attention for audio-visual fusion in dimensional emotion recognition," *IEEE J. Sel. Topics Signal Process.*, vol. 18, no. 3, pp. 444–458, Apr. 2024.
[10] S. Lee, D. K. Han, and H. Ko, "Multimodal emotion recognition fusion analysis adapting BERT with heterogeneous feature unification," *IEEE Access*, vol. 9, pp. 94557–94572, 2021.
[11] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, Curran Associates, Inc., 2017, pp. 5998–6008.

[12] J. Huang, J. Tao, B. Liu, Z. Lian, and M. Niu, "Multimodal transformer fusion for continuous emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 3507–3511.
[13] U. Bilotti, C. Bisogni, M. De Marsico, and S. Tramonte, "Multimodal emotion recognition via convolutional neural networks: Comparison of different strategies on two multimodal datasets," *Eng. Appl. Artif. Intell.*, vol. 130, 2024, Art. no. 107708.
[14] M. Gheini, X. Ren, and J. May, *Cross-Attention is All You Need: Adapting Pretrained Transformers for Machine Translation*. Stroudsburg PA USA: Association for Computational Linguistics, 2021.
[15] Y. Tang, Y. Hu, L. He, and H. Huang, "A bimodal network based on audio–text-interactional-attention with arcface loss for speech emotion recognition," *Speech Commun.*, vol. 143, pp. 21–32, 2022.
[16] Z. Liu, L. Cai, W. Yang, and J. Liu, "Sentiment analysis based on text information enhancement and multimodal feature fusion," *Pattern Recognit.*, vol. 156, 2024, Art. no. 110847.
[17] X.-C. Li, F. Zhang, Q. Hua, and C.-R. Dong, "A deep spatiotemporal interaction network for multimodal sentimental analysis and emotion recognition," *Inf. Sci.*, vol. 690, 2025, Art. no. 121515.
[18] M. K. Tellamekala, S. Amiriparian, B. W. Schuller, E. André, T. Giesbrecht, and M. Valstar, "COLD fusion: Calibrated and ordinal latent distribution fusion for uncertainty-aware multimodal emotion recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 2, pp. 805–822, Feb. 2024.
[19] F. Ma, S.-L. Huang, and L. Zhang, "An efficient approach for audio-visual emotion recognition with missing labels and missing modalities," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2021, pp. 1–6.
[20] J. Huang, J. Zhou, Z. Tang, J. Lin, and C. Y.-C. Chen, "TMBL: Transformer-based multimodal binding learning model for multimodal sentiment analysis," *Knowl.-Based Syst.*, vol. 285, 2024, Art. no. 111346.
[21] C. Doersch, "Tutorial on variational autoencoders," 2016, *arXiv:1606.05908*.
[22] S. Livingstone and F. Russo, "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*, vol. 13, no. 5, 2018, Art. no. e0196391.
[23] C. Busso et al., "Iemocap: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, pp. 335–359, 2008.
[24] D. Hazarika, S. Gorantla, S. Poria, and R. Zimmermann, "Self-attentive feature-level fusion for multimodal emotion detection," in *Proc. IEEE Conf. Multimedia Inf. Process. Retrieval*, 2018, pp. 196–201.
[25] X. Wang, M. Wang, H. Cui, and Y. Zhang, "A dual-channel multimodal sentiment analysis framework based on three-way decision," *Eng. Appl. Artif. Intell.*, vol. 137, 2024, Art. no. 109174.
[26] B. T. Atmaja and M. Akagi, "Multitask learning and multistage fusion for dimensional audiovisual emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process*, 2020, pp. 4482–4486.
[27] R. Fan, H. Liu, Y. Li, P. Guo, G. Wang, and T. Wang, "ATTA-NET: Attention aggregation network for audio-visual emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2024, pp. 8030–8034.
[28] M. Khan, J. Ahmad, W. Gueaieb, G. D. Masi, F. Karray, and A. E. Saddik, "Joint multi-scale multimodal transformer for emotion using consumer devices," *IEEE Trans. Consum. Electron.*, vol. 71, no. 1, pp. 1092–1101, Feb. 2025.
[29] M. Khan, P.-N. Tran, N. T. Pham, A. El Saddik, and A. Othmani, "Memocmt: Multimodal emotion recognition using cross-modal transformer-based feature fusion," *Sci. Rep.*, vol. 15, no. 1, Feb. 2025, Art. no. 5473.
[30] H. Zuo, R. Liu, J. Zhao, G. Gao, and H. Li, "Exploiting modality-invariant feature for robust multimodal emotion recognition with missing modalities," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 1–5.
[31] X. Zhao, X. Li, R. Jiang, and B. Tang, "Resolving multimodal ambiguity via knowledge-injection and ambiguity learning for multimodal sentiment analysis," *Inf. Fusion*, vol. 115, 2025, Art. no. 102745.
[32] C. Huang, J. Chen, Q. Huang, S. Wang, Y. Tu, and X. Huang, "AtCAF: Attention-based causality-aware fusion network for multimodal sentiment analysis," *Inf. Fusion*, vol. 114, 2025, Art. no. 102725.
[33] G. Q. Wang, J. Y. Li, Z. Wu, J. Xu, J. Shen, and W. Yang, "EfficientFace: An efficient deep network with feature enhancement for accurate face detection," *Multimedia Syst.*, vol. 29, pp. 2825–2839, 2023.
[34] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv: 2010.11929*.
[35] Y. Gong, Y.-A. Chung, and J. Glass, "Ast: Audio spectrogram transformer," 2021, *arXiv:2104.01778*.
[36] E. Ghaleb, M. Popa, and S. Asteriadis, "Multimodal and temporal perception of audio-visual cues for emotion recognition," in *Proc. 8th Int. Conf. Affect. Comput. Intell. Interact.*, 2019, pp. 552–558.
[37] H. R. Vaezi Joze, A. Shaban, M. L. Iuzzolino, and K. Koishida, "MMTM: Multimodal transfer module for CNN fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 13286–13296.
[38] E. Ghaleb, J. Niehues, and S. Asteriadis, "Multimodal attention-mechanism for temporal emotion recognition," in *Proc. IEEE Int. Conf. Image Process.*, 2020, pp. 251–255.
[39] C. Luna-Jiménez, D. Griol, Z. Callejas, R. Kleinlein, J. M. Montero, and F. Fernández-Martínez, "Multimodal emotion recognition on RAVDESS dataset using transfer learning," *Sensors*, vol. 21, no. 22, 2021, Art. no. 7665.
[40] S. Verbitskiy, V. Berikov, and V. Vyshegorodtsev, "ERANNs: Efficient residual audio neural networks for audio pattern recognition," *Pattern Recognit. Lett.*, vol. 161, pp. 38–44, 2022.
[41] P. Guo, Z. Chen, Y. Li, and H. Liu, "Audio-visual fusion network based on conformer for multimodal emotion recognition," in *Proc. Artif. Intell.: 2nd CAAI Int. Conf.*, 2022, pp. 315–326.
[42] K. Chumachenko, A. Iosifidis, and M. Gabbouj, "Self-attention fusion for audiovisual emotion recognition with incomplete data," in *Proc. 26th Int. Conf. Pattern Recognit.*, 2022, pp. 2822–2828.
[43] A. I. Middya, B. Nag, and S. Roy, "Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities," *Knowl.-Based Syst.*, vol. 244, 2022, Art. no. 108580.
[44] C. Luna-Jiménez, R. Kleinlein, D. Griol, Z. Callejas, J. M. Montero, and F. Fernández-Martínez, "A proposal for multimodal emotion recognition using aural transformers and action units on RAVDESS dataset," *Appl. Sci.*, vol. 12, no. 1, 2022, Art. no. 327.
[45] S. Li et al., "Multimodal emotion recognition in noisy environment based on progressive label revision," in *Proc. 31st ACM Int. Conf. Multimedia*, New York, NY, USA, 2023, pp. 9571–9575.
[46] Z. Li, L. He, J. Li, L. Wang, and W. Zhang, "Towards discriminative representations and unbiased predictions: Class-specific angular softmax for speech emotion recognition," in *Proc. Interspeech*, 2019, pp. 1696–1700.
[47] H. Xu, H. Zhang, K. Han, Y. Wang, Y. Peng, and X. Li, "Learning alignment for multimodal emotion recognition from speech," 2019, *arXiv: 1909.05645*.
[48] J. Liang, L. Ruichen, and Q. Jin, "Semi-supervised multi-modal emotion recognition with cross-modal distribution matching," 2020, *arXiv: 2009.02598*.
[49] P. Kumar, V. Kaushik, and B. Raman, "Towards the explainability of multimodal speech emotion recognition," in *Proc. Interspeech*, 2021, pp. 1748–1752.
[50] S. Padi, S. Sadjadi, D. Manocha, and R. Sriram, "Multimodal emotion recognition using transfer learning from speaker recognition and BERT-based models," in *Proc. Speaker Lang. Recognit. Workshop*, 2022, pp. 407–414.
[51] H. Zuo, R. Liu, J. Zhao, G. Gao, and H. Li, "Exploiting modality-invariant feature for robust multimodal emotion recognition with missing modalities," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 1–5.

Jiaming Zhang received the BS degree in mathematics from the Shenyang University of Technology, Shenyang, China in 2023. He is currently working toward the MEng degree in control science and engineering with the Same University. His research interests encompass Computer Game Theory, Audio Signal Processing, Acoustic Scene Classification, and Multimodal Emotion Recognition.

Zhijia Zhang received the BEng and MAEng degrees in mechanical engineering and automation from the Northeast University, Shenyang, China, in 1996 and 2002, respectively, and the PhD degree in pattern recognition and intelligent system from Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang, China, in 2006. He is currently a professor and the PhD Supervisor with the School of Artificial Intelligence, Shenyang University of Technology, Shenyang, China. His current research interests include product surface defect detection, safety monitoring of driver's driving behavior, and external environment awareness for autonomous driving.

Zhaojie Ju (Senior Member, IEEE) received the BS degree in automatic control, in 2005 and the MS degree in intelligent robotics from the Huazhong University of Science and Technology, China, in 2007, and the PhD degree in intelligent robotics from the University of Portsmouth, U.K, in 2010. He held research appointments with University College London, London, U.K., before he started his independent academic position with the University of Portsmouth. He is currently a full professor in machine learning and robotics and has authored or coauthored more than 300 publications in journals, book chapters, and conference proceedings. He has received nine Best Paper Awards and one Best AE Award in ICRA. His research interests include machine intelligence, pattern recognition and their applications in human robot interaction/collaboration, robot skill learning and healthcare & wearable robotics. Prof. Ju is an associate editor of several journals, such as *IEEE Transactions on Cybernetics*, *IEEE Transactions on Cognitive and Developmental Systems,* and *IEEE Transactions on neural networks and learning systems*.