

Identifying the Emotional State of Tourists in Rural Tourism Videos Using Multimodal Multi-Scale Attention Mechanisms

Zhiguo Yang

School of Economics and management, Handan University, Handan 056005, Hebei, China

Corresponding author: Zhiguo Yang (e-mail: 0yzg123@163.com)

ABSTRACT This study addresses the challenge of accurately identifying tourists' emotional states from unstructured rural tourism videos under diverse environmental stimuli by incorporating multimodal analysis of visual, acoustic, and textual information. To achieve robust and fine-grained emotion recognition, a Multimodal Multi-Scale Attention Transformer (MMSAT) framework is proposed for collaborative modeling of audiovisual and verbal signals. The framework employs an intra-modal attention mechanism to adaptively fuse multi-scale visual representations while utilizing a three-layer Transformer encoder to capture deep cross-modal interactions and complementary semantic information. The fused representations are subsequently used to predict emotional valence and arousal with high precision. Experimental results on a self-constructed rural tourism video dataset demonstrate that the proposed MMSAT model achieves Concordance Correlation Coefficient (CCC) scores of 0.823 and 0.805 for valence and arousal prediction, respectively, outperforming the strongest baseline model, Multimodal Transformer (MulT), by 0.042 and 0.053. The proposed framework provides an effective solution for objective emotion assessment in complex open environments and offers methodological insights into multimodal information propagation and intelligent signal processing, with potential implications for advanced electromagnetic information perception and cross-domain sensing applications.

INDEX TERMS multimodal emotion recognition, multiscale attention, cross-modal fusion, intelligent signal processing, electromagnetic information perception

I. INTRODUCTION

HIGH-quality development of rural tourism is a vital component of promoting rural revitalization and cultural tourism integration often achieved through the strategic revitalization of local heritage industries. A critical factor in achieving high-quality rural development is obtaining an objective and accurate assessment of tourist experiences. Historically, subjective evaluation methods for assessing the tourist experience, such as questionnaires and interviews, have intrinsic limitations (e.g., recall bias and social desirability bias) that hinder the effective evaluation of the dynamic and real-time emotional responses of tourists [1]. With the growing trend of utilizing video data in tourism contexts, emotion computing via computer vision provides an alternative method for quantifying the emotional experiences of tourists objectively [2,3]. However, the inherently complex natural environment, diverse tourist behaviors, and variable lighting and occlusion in rural tourism videos pose a significant challenge to automated emotion recognition technology [4]. Achieving fine-grained and highly robust

emotion state recognition in such uncontrolled environments has become a cutting-edge interdisciplinary topic linking information technology and tourism management disciplines. To address the limitations of single-modality approaches, research on multimodal emotion recognition has emerged. Early studies often employed simple fusion strategies, such as feature-level concatenation or decision-level weighting, which improved performance to some extent but neglected the deep interactions between modalities [5,6]. To model more complex cross-modal relationships, tensor fusion networks explicitly capture high-order interactions between modalities through outer product operations, but their computational complexity is high, and they are prone to overfitting [7]. Subsequently, fusion methods based on the low-rank assumption and modality-invariant/specific representation learning models were optimized from the perspectives of computational efficiency and representation decoupling, respectively [8]. In recent years, the Transformer architecture [9,10] has shown significant advantages in multimodal sequence modeling due to its powerful self-attention and cross-attention mechanisms,

becoming the current mainstream technical path. Despite significant progress in multimodal fusion technology, its application in the specific scenario of rural tourism still faces many unresolved challenges. First, existing methods focus primarily on intermodal fusion while neglecting the hierarchical nature of features within a single modality [11,12]. Tourist emotions are characterized by macroscopic scene context, mesoscopic body language, and microscopic facial expressions, but current research lacks collaborative modeling of these multi-scale features [13]. Second, in complex outdoor environments, audio and visual signals are often contaminated by noise, and most models lack the ability to dynamically perceive and adapt to modal reliability, leading to a sharp decline in performance in scenarios with low signal-to-noise ratios [14]. Furthermore, current research is mostly focused on controlled laboratory environments or standard datasets. For complex and unstructured video data in real rural tourism scenarios, the generalization ability and robustness of existing models still need to be systematically verified [15]. Previous research has significant shortcomings in the scale adaptability of feature extraction, the dynamic robustness of the fusion process, and the generalization ability for real-world scenarios. To systematically address the aforementioned challenges, this paper proposes a Multimodal Multi-Scale Attention Transformer (MMSAT) model for rural tourism videos. The core innovations of this research lie in three aspects: First, at the feature representation level, addressing the hierarchical nature of emotional signals and the problem of single-scale approaches in existing methods, this paper innovatively constructs a “macro-meso-micro” visual multi-scale extraction pathway. By separating the global scene, human pose, and facial expression regions through object detection, and applying an intra-modal attention mechanism, dynamic adjustment and adaptive fusion of visual cues at different scales are achieved. In addition, at the information fusion level, to solve the problems of inflexible and difficult-to-construct models for dealing with complex environmental interferences in traditional fusion methods, this paper utilizes a Transformer-based cross-modal interaction framework that uses cross-attention as the main data processing mechanism, where features from each modality are used as queries for the retrieval of the most relevant evidence from other modalities for deep semantic alignment and complementarity. Furthermore, at the paradigm validation level, this paper provides a dedicated rural tourism video dataset, validating the model’s generalization ability and practical application, especially in accurately measuring affective responses to cultural stimuli. This yields a high-performing technical resource and a new perspective for understanding the theoretical development of multimodal emotion computing in complex open environments based on multi-scale representation learning and dynamic attention fusion, providing a solid scientific basis for connecting state-of-the-art artificial intelligence technologies to the practice of tourism management.

II. METHODS

Overall Framework

To address the challenges of emotion recognition in rural tourism videos, this paper proposes a MMSAT. It constructs a “macro-meso-micro” visual pathway to achieve hierarchical feature extraction and applies intra-modal and cross-modal attention, dynamically adjusting and fusing multi-scale and multimodal information. The overall architecture of the model is shown in Figure 1, comprising four core stages: multimodal multi-scale feature extraction, intra-modal attention fusion, Transformer-based intermodal fusion, and emotion regression.

As shown in Figure 1, the model first performs multimodal multi-scale feature extraction. For the visual modality, the original video frames are processed by an object detector and deconstructed into image regions at three scales: macro (global scene), meso (human body bounding box), and micro (facial bounding box). Deep feature vectors are then extracted from each region using a convolutional neural network (CNN) backbone. Simultaneously, the original audio waveform is converted into a Mel spectrogram, and its temporal structure features are captured using a temporal CNN. The text transcribed using automatic speech recognition (ASR) technology is input into a pre-trained Bidirectional Encoder Representations from Transformers (BERT) model to obtain semantically rich text embedding vectors. This stage transforms the original video data into a set of structured, semantically rich deep features. Subsequently, the architecture enters the hierarchical attention fusion stage. The model does not treat all features equally, but first filters them within the same modality: the three scale features of the visual modality pass through an intra-modal attention layer, which automatically learns and assigns weights through a feedforward neural network, dynamically fusing them into a weighted “visual essence feature vector”, thus achieving primary information purification within the modality. Afterward, the essential feature vectors from visual, audio, and text modalities are concatenated into a sequence and input into a Transformer encoder. At this stage, the intermodal attention mechanism begins to function; its self-attention operation allows features from each modality to serve as queries, retrieving and focusing on the most relevant information from other modalities, thereby achieving deep, context-aware cross-modal feature collaboration and complementarity. Finally, the [CLS] token output by the Transformer, as an aggregated representation of the entire input sequence, is fed into a fully connected layer to regress continuous sentiment valence and arousal values.

Multimodal Data Preprocessing and Feature Extraction

Visual Modal Multi-Scale Feature Extraction

To accurately capture multi-layered emotional information, from scene atmosphere to individual facial expressions,

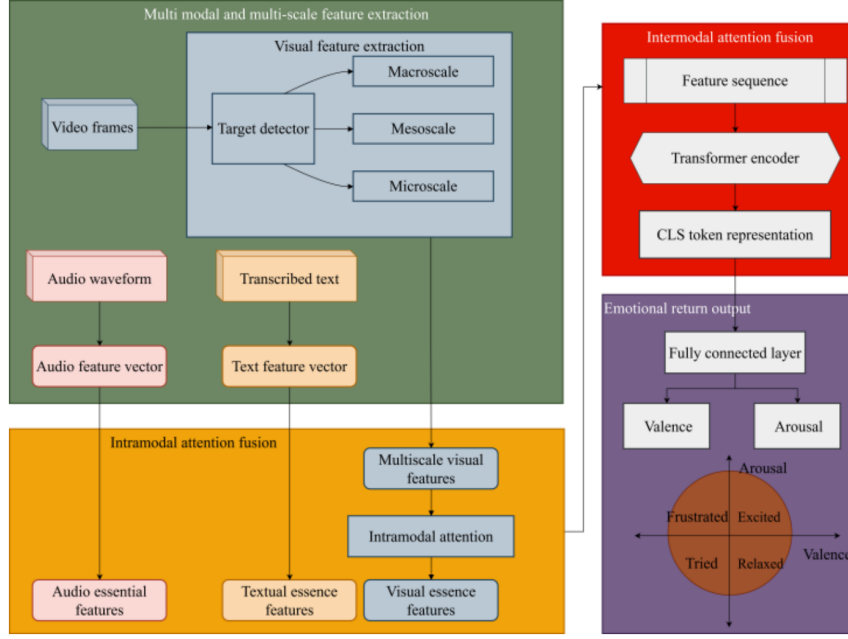


FIGURE 1. Overall model architecture

this paper constructs a three-layer visual feature extraction pipeline, the process of which is shown in Figure 2.

Visual feature extraction begins with the object detection stage. First, the raw video frames are fed into a pre-trained YOLOv5 object detector [16,17] to simultaneously identify human bodies and faces in the image. Based on the detection results, the raw frames are divided into three levels of regions of interest: a macroscopic global scene, a mesoscopic maximum human body bounding box, and a microscopic corresponding face bounding box. Each region image is cropped and scaled to a uniform size to serve as input for the subsequent feature extraction network. To ensure the stability of facial features, keypoint-based alignment preprocessing is applied to the face region. Given a video frame I , an object detector produces three levels of regions: a macroscopic region I_g , a mesoscopic region I_b , and a microscopic region I_f . The images of these three regions are then input into a shared-weights ResNet-50 backbone network pre-trained on ImageNet to extract deep features. The formulas for this process are given as follows:

$$\begin{aligned} F_g &= \text{ResNet}(I_g), & F_b &= \text{ResNet}(I_b), \\ F_f &= \text{ResNet}(I_f) \end{aligned} \quad (1)$$

$$F_{\text{visual}} = [F_g, F_b, F_f] \quad (2)$$

Audio Modal Feature Extraction

The mono audio stream is separated from the video and resampled to a uniform sampling rate of 16 kHz to normalize the input. For the audio signal $x[n]$, feature extraction begins with the calculation of the Mel spectrogram, which simulates the nonlinear perception of frequency by the human ear [18,19]. A first-order high-pass filter

$H(z) = 1 - \alpha z^{-1}$ ($\alpha = 0.97$) is applied to the signal to enhance high-frequency components. The signal is divided into frames with a frame length of 25 ms and a frame shift of 10 ms, resulting in a series of overlapping short-time frames $x_i[m]$. A Hamming window is added to each frame $x_i[m]$, and a short-time Fourier transform is performed. The square of the modulus is then taken to obtain the short-time power spectrum $P_i[k]$, where the magnitude-squared operation corresponds to the short-time power spectrum.

$$P_i[k] = \left| \sum_{m=0}^{N-1} x_i[m] w[m] e^{-j2\pi km/N} \right|^2 \quad (3)$$

In Formula 3, $w[m]$ denotes the Hamming window, and N represents the number of points in the fast Fourier transform (FFT). The power spectrum $P_i[k]$ is filtered through a set of 40 triangular Mel-scale filters $\{H_m[k]\}_{m=1}^{40}$, and the logarithm of the output energy of each filter is taken to obtain the log-Mel spectrum:

$$S[i, m] = \ln \left(\sum_k P_i[k] \cdot H_m[k] \right) \quad (4)$$

An audio segment of length T is transformed into a time-frequency representation $S \in R^{T \times 40}$, that is, a Mel spectrogram [20,21]. This log-Mel spectrogram is input into a temporal CNN for deep feature learning. The network consists of three convolutional blocks, each containing a one-dimensional convolutional layer, an activation function, and a max-pooling layer. The one-dimensional convolution is performed along the time axis, and its core operation is formulated as follows:

$$h_l[t] = \text{ReLU}(W_l * h_{l-1}[t] + b_l) \quad (5)$$

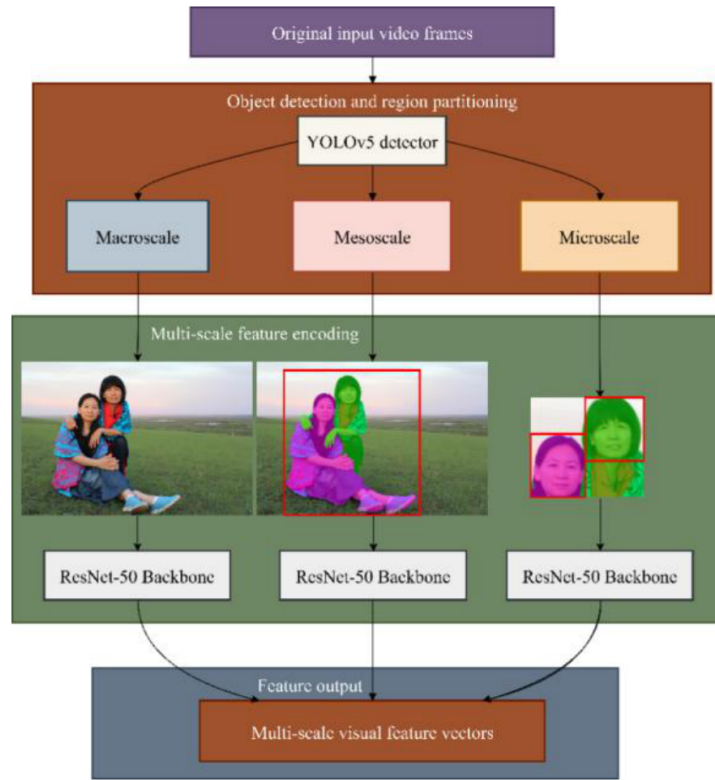


FIGURE 2. Visual multi-scale feature extraction

In Formula 5, $*$ represents a one-dimensional convolution operation; W_l and b_l are the convolution kernel and bias of the l -th layer; h_{l-1} is the feature map of the previous layer. The width of the convolution kernel is designed to cover multiple adjacent frames, thereby capturing local temporal patterns. The output feature map is then processed by a global average pooling layer to aggregate into a fixed-dimensional feature vector $F_a \in R^{d_a}$, which serves as the final representation of the audio modality. The formula is:

$$F_a = \text{GlobalAveragePooling}(h_L) \quad (6)$$

In Formula 6, L is the final convolutional layer. This vector F_a encapsulates key temporal patterns related to emotion in the audio signal and can be fed into the subsequent cross-modal fusion module.

Text Modal Feature Extraction

Textual modalities carry tourists' direct emotional expressions and semantic content, providing high-level, interpretable cues for emotion recognition. To extract and utilize textual semantic information from raw videos, this work constructs a text processing pipeline consisting of ASR and a pre-trained language model, converting spoken language into high-quality text semantic embeddings. The separated audio stream is transcribed into the original text sequence using ASR technology. The ASR system utilized is the open-source Whisper medium model. To assess its reliability in the target domain, the Word Error Rate (WER) was evaluated on a held-out subset of the rural tourism dataset,

yielding a WER of approximately 15%. While this indicates potential noise in the textual stream, the subsequent BERT model is leveraged for its contextual understanding and robustness to minor transcription errors. Nevertheless, the performance drop observed when removing the text modality underscores its contributory role despite potential ASR inaccuracies. Given audio features, the ASR system aims to find the most likely word sequence $W = \{w_1, w_2, \dots, w_N\}$:

$$\hat{W} = \arg \max_W P(W | S) \quad (7)$$

In Formula 7, S represents the acoustic features of the audio. This study directly models $P(W | S)$ to generate accurate transcribed text $\hat{W}$ with punctuation. The transcribed text $\hat{W}$ is fed into a pre-trained BERT model to extract deep context-sensitive word vectors. BERT uses a Transformer encoder architecture, encoding the input text through a multi-layer self-attention mechanism. The text is first tokenized into sub-word units, and special [CLS] tokens and sentence separators [SEP] are added to form the input sequence CLS. The model processes this sequence through multiple layers of Transformer blocks; the operations of each layer can be summarized as follows:

$$H_l = \text{TransformerBlock}(H_{l-1}) \quad (8)$$

In Formula 8, H_l is the sequence of hidden states of all tokens in the l -th layer, and H_0 is the input embedding. The core computation of the self-attention mechanism is as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (9)$$

In Formula 9, Q, K, V are the query, key, and value matrices obtained by linear projection of the input H, respectively, and d_k is the dimension of the key vector. This paper takes the hidden state $h_{[\text{CLS}]}^L \in R^{d_h}$ of the [CLS] token output by the last Transformer layer as the aggregate semantic representation of the entire input text sequence:

$$F_t = h_{[\text{CLS}]}^L \quad (10)$$

Text feature vectors, along with visual and audio features, are input into the subsequent cross-modal fusion module, providing crucial semantic-level evidence for emotion recognition.

Intramodal Multiscale Attention Fusion

Visual modalities comprise feature vectors $F_g, F_b, F_f \in R^d$ at three scales: macro, meso, and micro. However, the importance of these scales is not constant across different emotional states and contextual situations. To address this dynamic importance issue, this paper designs an intra-modal multi-scale attention fusion module, enabling the model to automatically learn and assign optimal weights to visual features at different scales, thereby achieving intra-modal information purification and adaptive fusion. The input to this module is a set of features at the three scales of the visual modality $F = \{F_g, F_b, F_f\}$. Its core is a content-based multi-head attention mechanism. Each feature vector $F_i \in R^d$ (where $i \in \{g, b, f\}$) is linearly projected onto the query, key, and value spaces:

$$Q_i = W_Q F_i, \quad K_i = W_K F_i, \quad V_i = W_V F_i \quad (11)$$

In Formula 11, $W_Q, W_K, W_V \in R^{d \times d_k}$ are the learnable weight matrices. To capture information from different subspaces, multi-head attention is employed. For the j -th head, different projection matrices W_j^Q, W_j^K, W_j^V are used to obtain the corresponding $Q_{i,j}, K_{i,j}, V_{i,j}$. The attention score for each head is calculated using a scaled dot product. For each scale i , the formula for its attention weights $\alpha_{i,k}^j$ with all other scales (including itself) is:

$$\alpha_{i,k}^j = \frac{\exp(\langle Q_{i,j}, K_{k,j} \rangle / \sqrt{d_k})}{\sum_{m \in \{g,b,f\}} \exp(\langle Q_{i,j}, K_{m,j} \rangle / \sqrt{d_k})} \quad (12)$$

In Formula 12, $\langle \cdot, \cdot \rangle$ represents the dot product operation. Afterward, the output of each head is a weighted sum of all value vectors:

$$\text{head}_{i,j} = \sum_{k \in \{g,b,f\}} \alpha_{i,k}^j V_{k,j} \quad (13)$$

The outputs of all heads are concatenated and projected through a linear layer W_O to obtain refined features F'_i at that scale:

$$F'_i = \text{LayerNorm}(F_i + \text{Dropout}(W_O[\text{head}_{i,1}; \dots; \text{head}_{i,H}])) \quad (14)$$

After further processing by a feedforward network, three refined scale features F'_g, F'_b, F'_f are obtained. To fuse them into a unified visual representation, a learnable context vector $c \in R^{d_k}$ is applied to compute the global importance weight β_i of each refined feature:

$$e_i = c^T \tanh(W_a F'_i + b_a), \quad \beta_i = \frac{\exp(e_i)}{\sum_{j \in \{g,b,f\}} \exp(e_j)} \quad (15)$$

In Formula 15, W_a and b_a are learnable parameters. The final visual essence feature vector $F_{\text{visual}} \in R^d$ is a weighted sum of the three refined features:

$$F_{\text{visual}} = \sum_{i \in \{g,b,f\}} \beta_i F'_i \quad (16)$$

Through this intra-modal attention fusion module, the model can dynamically and selectively emphasize the visual scales most relevant to the current emotional context, while suppressing scales that may be affected by noise or have low information content, thereby generating a more discriminative visual representation and laying a solid foundation for subsequent cross-modal fusion.

Transformer-Based Intermodal Attention Fusion

After obtaining the refined features of each modality, to achieve deep cross-modal interaction, this paper adopts the Transformer encoder [22,23] as the core architecture for inter-modal fusion. Its core parameter configuration is shown in Table 1.

TABLE 1. Core parameters of the Transformer encoder

Parameter name	Value	Instructions
Encoder layers	3	Number of stacked Transformer blocks
Number of attention heads	8	Number of heads in the multi-head attention mechanism
Hidden layer dimension	512	Dimension of input/output features and internal main path
Feedforward network dimension	2048	Internal dimensions of feedforward networks in each Transformer block
Attention dimension	64	Dimension of query/key/value vectors in each attention head
Dropout	0.1	Dropout ratio used to prevent overfitting

The input to the fusion process is a sequence $Z_0 = [F_{\text{visual}}; F_{\text{audio}}; F_{\text{text}}] \in R^{3 \times d}$, consisting of visual, audio, and textual essence features, where $d = 512$. This sequence is fed into a Transformer encoder consisting of $N = 3$ identical stacked layers. In each layer l , cross-modal interaction is achieved through a multi-head self-attention mechanism. First, the input Z_{l-1} is linearly projected into a query, key, and value matrix:

$$\begin{aligned} Q_l^h &= Z_{l-1} W_l^{Q,h}, \\ K_l^h &= Z_{l-1} W_l^{K,h}, \\ V_l^h &= Z_{l-1} W_l^{V,h} \end{aligned} \quad (17)$$

In Formula 17, $h \in [1, H]$ represents the attention head index, $W_l^{Q,h}, W_l^{K,h} \in R^{d \times d_k}$, and $W_l^{V,h} \in R^{d \times d_v}$ are learnable parameters, and $d_k = d_v = 64$. Scaled dot product attention is calculated for each head:

$$\text{head}_l^h = \text{Softmax} \left(\frac{Q_l^h (K_l^h)^T}{\sqrt{d_k}} \right) V_l^h \quad (18)$$

The outputs of all heads are concatenated and projected:

$$\text{MHAttention}(Z_{l-1}) = [\text{head}_l^1; \dots; \text{head}_l^H] W_l^O \quad (19)$$

In Formula 19, $W_l^O \in R^{H \cdot d_v \times d}$. Afterward, the attention output goes through a residual connection and layer normalization:

$$U_l = \text{LayerNorm}(Z_{l-1} + \text{Dropout}(\text{MHAttention}(Z_{l-1}))) \quad (20)$$

Next, U_l passes through a positional feedforward network, which consists of two linear transformations and an intermediate activation function:

$$\text{FFN}(U_l) = \text{Dropout}(W_l^2 \cdot \text{ReLU}(W_l^1 U_l + b_l^1) + b_l^2) \quad (21)$$

In Formula 21, $U_l \in R^d$, $W_l^1 \in R^{d \times d_{ff}}$, $W_l^2 \in R^{d_{ff} \times d}$, and $d_{ff} = 2048$. Similarly, applying residual connections and layer normalization yields the final output of this layer:

$$Z_l = \text{LayerNorm}(U_l + \text{Dropout}(\text{FFN}(U_l))) \quad (22)$$

After N layers of such processing, the features at each position in the sequence have been enriched and contextualized by features from other modalities. This paper takes the feature at the first position of the final output sequence Z_N as the fused multimodal joint representation $H_{\text{fused}} \in R^d$:

$$H_{\text{fused}} = Z_N[0] \quad (23)$$

This vector H_{fused} encapsulates complementary information from all modalities through deep interaction and can be fed into the final regression layer for sentiment dimension prediction.

Output Layer and Loss Function

To map the fused multimodal joint representation to a continuous dimensional emotion space, this paper employs a simple fully connected layer as the output layer. The input to this layer is the [CLS] token vector $H_{\text{fused}} \in R^{512}$ from the Transformer encoder, and its output is a two-dimensional vector $\hat{y} \in R^2$, corresponding to the predicted values of emotional valence and arousal, respectively, as shown in the formula:

$$\hat{y} = [\hat{v}, \hat{a}] = W_o H_{\text{fused}} + b_o \quad (24)$$

In Formula 24, $W_o \in R^{2 \times 512}$ and $b_o \in R^2$ are the learnable weight matrix and bias vector, respectively. Different regions in the dimensional emotion space [24,25] are associated with discrete descriptions of emotional states, as shown in Table 2.

TABLE 2. Emotional descriptions of typical regions in valence–arousal (VA) two-dimensional space

Valence (V)	Arousal (A)	Typical emotional descriptors
High	High	Excited, Awed, Delighted
High	Low	Relaxed, Content, Peaceful
Low	High	Frustrated, Anxious, Irritated
Low	Low	Bored, Tired, Disappointed
Neutral	Neutral	Neutral, Calm

Table 2 provides the sentiment semantic mappings for typical quadrants in the VA model [26,27], offering a conceptual framework for subsequent analysis. Given that sentiment prediction is a regression task, this paper uses mean squared error (MSE) loss as the optimization objective of the model. This loss function measures the difference between the model's predicted values $\hat{y}_i = [\hat{v}_i, \hat{a}_i]$ and the true labeled values $y_i = [v_i, a_i]$. For a training batch containing N samples, the loss L is calculated as follows:

$$L = \frac{1}{N} \sum_{i=1}^N [(v_i - \hat{v}_i)^2 + (a_i - \hat{a}_i)^2] \quad (25)$$

III. EXPERIMENTS

Dataset Construction

To evaluate the effectiveness of the proposed model, this study independently constructs a dedicated rural tourism video dataset. Data collection is conducted at several representative rural tourist attractions, using a smartphone (iPhone 14 Pro) and a portable action camera (DJI Osmo Action 3) under natural light to ensure high-quality video (1080p resolution, 30 FPS) and audio (48 kHz sampling rate). The recording scenes cover a variety of typical scenarios, including rural scenery viewing, agricultural experience activities, participation in traditional folk customs, and rural homestay leisure, to comprehensively capture tourists' emotional responses in different contexts.

Table 3 provides the main statistics for this dataset, which collects data from 89 participants who produced 1,285 videos. From these videos, 15,420 valid segments are extracted for cross-modal video emotion (CVE) analysis. Each segment is labeled for valence and arousal using a continuous scale from 0–1 by five trained raters independently. The final ground-truth label for each segment was determined by taking the median of the five independent ratings. The average valence for the dataset is 0.68, which is above the neutral value of 0.5, indicating that, on average,

experiences of rural tourism are usually positive based on expectations about the types of experiences that can result in positive emotions. The dataset also contains standard deviations, which present an additional challenge when developing models.

TABLE 3. Statistical information of rural tourism video dataset

Statistical project	Value/Description
Total number of videos	1,285
Total duration	About 42.8 hours
Number of participants	89
Scene distribution	Rural scenery viewing (42%), agricultural experience activities (28%), traditional folk customs participation (19%), rural homestay leisure (11%)
Number of annotated segments	15,420
Average duration of fragments	10 s
Valence mean (standard deviation)	0.68 (0.22)
Arousal mean (standard deviation)	0.59 (0.25)

To ensure a fair and generalizable evaluation, the dataset was partitioned following a speaker-independent protocol. The 89 participants were randomly split into training, validation, and test sets in a 7:1:2 ratio. Correspondingly, the video segments were divided, resulting in 10,794 segments for training, 1,542 for validation, and 3,084 for testing. This partitioning ensures that the model is evaluated on completely unseen individuals, providing a robust test of its generalization ability to new tourists. All baseline models and the proposed model were trained and tested using this identical data split to guarantee a fair comparison. This study was approved by the Ethics Committee of Handan University and was performed in accordance with the principles of the Declaration of Helsinki. All eligible participants signed an informed consent form.

A. EXPERIMENTAL SETUP

This experiment is implemented using the PyTorch 2.0 deep learning framework. The model employs the AdamW optimizer with an initial learning rate of 3×10^{-5} and a cosine annealing scheduling strategy. The weight decay coefficient is set to 0.01, and the training batch size is 32. The model is trained for 100 epochs on four NVIDIA RTX 4090 GPUs. To improve training stability, all input features are layer-normalized, and dropout with a ratio of 0.1 is applied before each fully connected layer. To comprehensively evaluate model performance, three complementary regression metrics are used. MSE measures the squared deviation between predicted and true values and is sensitive to outliers; the formula is:

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (26)$$

Mean Absolute Error (MAE) provides an intuitive measure of error, with the formula:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i| \quad (27)$$

Concordance Correlation Coefficient (CCC) [28,29] considers both relevance and accuracy and is the most crucial metric in dimensional emotion recognition tasks. The formula is:

$$\text{CCC} = \frac{2\rho\sigma_y\sigma_{\hat{y}}}{\sigma_y^2 + \sigma_{\hat{y}}^2 + (\mu_y - \mu_{\hat{y}})^2} \quad (28)$$

In Formula 28, ρ is the Pearson correlation coefficient, and μ and σ are the mean and standard deviation, respectively. To verify the advancement of the proposed model (MMSAT), it is compared with six advanced multimodal fusion models representing different technical approaches: Tensor Fusion Network [30,31] (TFN, explicit high-order interaction based on tensor outer product), Low-Rank Fusion Network [32,33] (LFM, efficient implementation of low-rank approximation of TFN), Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis (MISA, decoupling features into modality-invariant and modality-specific components), Multimodal Transformer [34,35] (MulT, a benchmark for cross-modal fusion based on Transformer cross-attention), Self-Supervised Multi-Task Multimodal (Self-MM, using multi-task self-supervised learning to learn robust single-modal representations), and Quality-Aware Multimodal Fusion (QMF, a fusion method that explicitly evaluates modal quality). These baselines cover the mainstream technological paradigms of current multimodal emotion recognition, ensuring the comprehensiveness and fairness of the comparison.

IV. RESULTS

Tourist Emotional State Prediction Results

To evaluate the overall performance of the MMSAT (proposed) model, this study conducts a comprehensive comparison with six state-of-the-art baseline models on the test set. The prediction results for valence and arousal are shown in Figure 3.

The MMSAT model demonstrates superior performance in the datasets used for valence and arousal predictions. In terms of the main metric measured through the CCC, MMSAT surpasses the baseline model (MulT) by 0.042 for valence prediction and 0.053 for arousal prediction, with MMSAT exhibiting the least amount of error (lowest MAE/MSE) compared to the baseline MulT model. The two main strengths of the MMSAT model are outlined below. First, multi-scale visual feature extraction and intra-modal attention mechanisms allow MMSAT to flexibly concentrate on key visual cues across multiple levels of visual presentation (including face and scene context), providing more complete visual representations with better performance compared to single-scale or simple splicing. Second, using transformer-based intermodal fusion allows MMSAT to

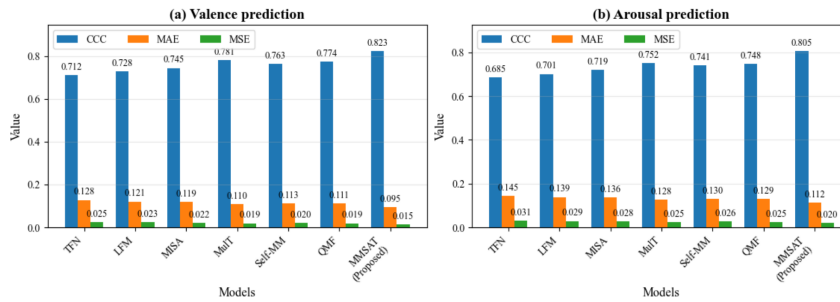


FIGURE 3. Emotional state prediction results: (a) Valence prediction results; (b) Arousal prediction results

dynamically model nonlinear interactions across modalities and combine them, enabling MMSAT to better incorporate complementary information than any other method of fusing modalities, including tensor fusion and early/late modal fusion. Therefore, MMSAT exhibits more robust emotion recognition in complex rural tourism environments than any other current method.

Ablation Experiment

To verify the effectiveness of each core component in the MMSAT model, ablation experiments are systematically conducted, and the results are shown in Table 4. Removing any key module results in a significant decrease in model performance, demonstrating the necessity of the overall architecture design presented in this paper.

TABLE 4. Comparison of ablation test performance (evaluation index is CCC)

Model variants	Valence (CCC)	Arousal (CCC)	ΔV	ΔA
Complete model (MMSAT)	0.823	0.805	–	–
–Intra-modal attention	0.791	0.773	–0.032	–0.032
–Modal Transformer	0.768	0.749	–0.055	–0.056
–Text modality	0.801	0.789	–0.022	–0.016
–Multi-scale visual features	0.780	0.762	–0.043	–0.043

Note: ΔV and ΔA represent the decrease in CCC relative to the full model on valence and arousal. The “intermodal Transformer” variant uses feature concatenation followed by a fully connected layer for fusion. The findings from the ablation studies quantifying the role of each component on model performance show clearly that model performance suffers the most when the intermodal Transformer is removed from the model: the valence/arousal CCC drops by 0.055/0.056, respectively. This indicates that combining two modalities through simple concatenation and fusion cannot represent the complex nonlinear interaction between these modalities, thus offering further justification for the need to use a Transformer-based representation for deep cross-modal modeling. The removal of multi-scale visual features also results in significant drops in model performance: both valence/arousal CCC drop by 0.043 with the removal of these features. This illustrates the importance of modeling the global scene, human pose, and facial expressions separately and collectively, and how the model cannot account

for variations in emotional carriers across different contexts without multi-scale visual feature inputs. There are also significant performance drops when intra-modal attention is removed: both valence/arousal CCC drop by 0.032. The inclusion of multi-scale visual features is not enough to overcome the loss of a dynamic weighting mechanism for these features. When intra-modal attention is not employed for the fusion of visual data, there is less accuracy in the generated visual representations. Removing the text modality only has a small effect on performance (valence/arousal CCC drop by 0.022 and 0.016, respectively). This is expected for certain emotions, such as joy and surprise, especially in a rural tourism context. In these cases, strong visual and audio signals convey enough of the emotional meaning without needing the additional information from the text modality. In conclusion, all components together constitute a powerful and robust multimodal emotion recognition system. While the most significant performance drop occurred upon removing the intermodal Transformer, highlighting the critical role of deep cross-modal interaction, it is important to emphasize that the performance degradation resulting from removing the core innovation—multi-scale visual feature extraction with intra-modal attention fusion (0.043/0.043)—is only marginally smaller. This demonstrates the substantial effectiveness and necessity of the multi-scale features. The Transformer, as a powerful fusion architecture, relies heavily on the quality of its input features for optimal performance. The proposed innovation provides precisely this: a more discriminative and information-rich visual representation, purified and adaptively weighted within the modality, which forms a crucial foundation for the overall performance gain. Therefore, the ablation results should be interpreted as evidence that the proposed multi-scale feature extraction and intra-modal attention mechanism work in effective synergy with the cross-modal attention fusion architecture to jointly achieve the model’s superior performance. While the ablation study demonstrates the collective importance of multi-scale features, it is acknowledged that a more granular comparison against traditional single-scale baselines (e.g., using only macro-scale or only micro-scale features) would further solidify the necessity of the proposed three-scale design. Future work will include such direct comparisons to explicitly quantify the advantage of hierarchical feature extraction over conventional single-scale approaches.

Attention-Based Analysis on Emotional Fragments

To explore the interpretability of the internal decision-making mechanism of the MMSAT model, the dynamic changes of intra-modal and intermodal attention weights under different emotional segments are analyzed. Figure 4 shows the evolution of the three-scale weight allocation within the visual modality and the cross-modal attention score, revealing the model's ability to adaptively focus on key cues.

The ability of the model to adaptively shift its visual focus depending upon the type of emotion is shown clearly in Figure 4(a). For example, high-arousal emotions such as Excited (mesoscale weight 0.50), Frustrated (mesoscale 0.45), and Anxious (mesoscale 0.40) all place significant importance on mesoscale (human posture) features; this is because, during the presentation of these high-arousal emotions, large-scale body movements are used as the most direct and clear external expression of those emotions. In contrast, low-arousal emotions such as Relaxed (macroscale 0.55), Peaceful (macroscale 0.60), and Bored (macroscale 0.65) rely primarily on the macroscale (scene context) for cues; calm emotional states are therefore more fully defined by the overall context of their environment. Furthermore, there is a distinctly greater contribution to the model by the microscale (face) feature for the emotions of Delighted (microscale 0.35) and Disappointed (microscale 0.35), demonstrating that the model is capable of capturing more subtle emotional expressions. Through adaptive weight allocation, the intra-modal attention mechanism effectively integrates multi-scale visual information. Multimodal data can be merged, and a neural network can be trained to identify the most reliable data for a particular emotional context, similar to how a human would do this, as shown in Figure 4(b). The visual modality consistently receives the majority of the attention weight (normally above 0.45). This is also supported by psychological research, which states that the perception of another's emotions is primarily mediated through their visual presentation. When there are instances of high levels of negative arousal, like Frustrated (0.40) or Anxious (0.45), the audio (tone of voice) modality is more heavily weighted than the visual modality. This can be explained by the fact that, in everyday interactions, when people are frustrated or anxious, their tone of voice (complaints, sighs) often expresses their true emotional state more than their facial expression, which may be misrepresented or hidden. The model has successfully captured this phenomenon. In addition, the audio and text modalities are equal (0.30) for the Disappointed segment, indicating that the model appears to seek support through semantic context (i.e., genuine disappointment) for the emotional context. Overall, this indicates that the model can dynamically evaluate how to weight different modalities based on the most valid source of information in a specific context.

Visualization of Sentiment Prediction Results

To visually compare the prediction accuracy of each

model over time, this paper selects a representative video clip containing 50 consecutive time points, which fully demonstrates the dynamic change of tourists' emotions from Calm to Excited. Figure 5 shows a comparison between the prediction curves of each model and the actual values in the valence and arousal dimensions.

All models generally seem able to predict the overall movement of emotion transitioning from Calm to Excited from a valence point of view, as can be seen in the figure below, but there are large discrepancies in their accuracy of prediction and the timing of when those predictions are correct (timeliness). The baseline model has a pronounced lag effect when predicting the transitional phase of emotions to Excited and exhibits significant smoothing effects, indicating that this model is not able to track rapid changes in the actual value as effectively as the MMSAT model. The MMSAT model's predictions are the closest to the actual values, not only exhibiting the general upward trend of the transitional process of emotion but also capturing the variations in emotions that occur between the two states (between Calm and Excited). This is due to the MMSAT model's multi-scale visual feature extraction mechanism, which allows the model to simultaneously gather different, yet complementary, emotional data from facial expression cues, body movement cues, and the atmosphere created by the scene. In doing so, this model can provide an accurate representation of the overall changes in emotional states. When analyzing and comparing the performance of the models based on the arousal dimension (see Figure 5b), it becomes increasingly clear that the differences between model performances in this dimension are even larger than those in the valence dimension. For example, the traditional models exhibit substantial prediction bias during the peak segments of emotional arousal, or frames 30 to 40. This indicates that traditional models are not very good at recognizing high-arousal emotions (i.e., recognizing emotions that are highly excited). The MMSAT model's predictions closely match the true peak values in this area due to its cross-modal attention mechanism. Specifically, this cross-modal attention mechanism integrates audio, visual, and text data in real time and automatically increases the attention weights to those modalities when an excited state is detected. As a result, the MMSAT model can accurately identify high-arousal emotions. In addition, the MMSAT model continues to provide good tracking across the full time series and, therefore, exhibits consistent performance across different levels of emotional intensity. To assess the consistency between the predicted and actual values in the overall distribution of the emotion space, a scatter plot of the actual values and model predictions is drawn in the VA two-dimensional space, as shown in Figure 6.

Figure 6 shows the differences in how each of the three labels distributes within the VA emotion space. The actual (solid circle data) represents an ideal distribution based on emotion theory, as they are all located within the four quadrants (i.e., strong positive, weak positive, weak nega-

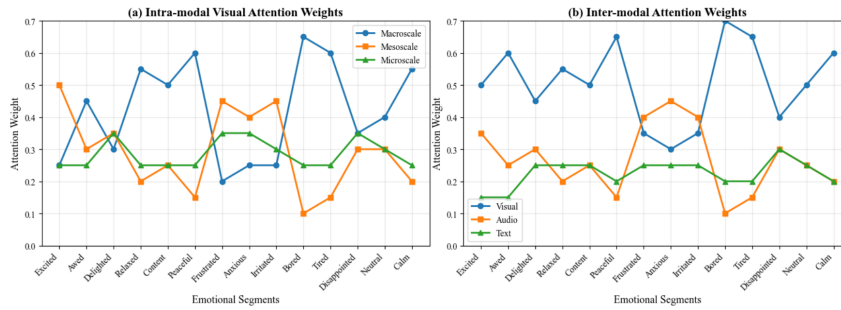


FIGURE 4. Attention analysis based on emotion segments: (a) Multi-scale attention weights within the visual modality; (b) Intermodal attention weights

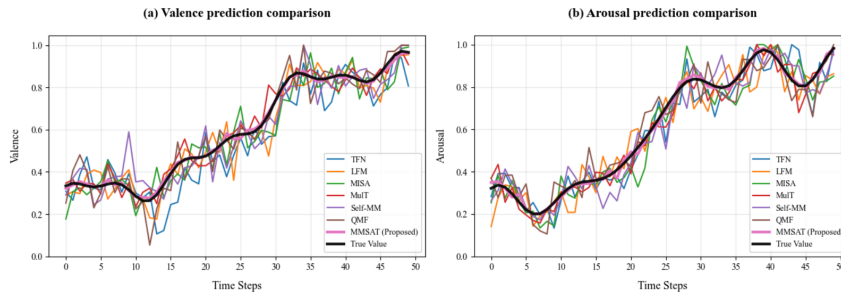


FIGURE 5. Visualization of sentiment prediction: (a) Valence; (b) Arousal

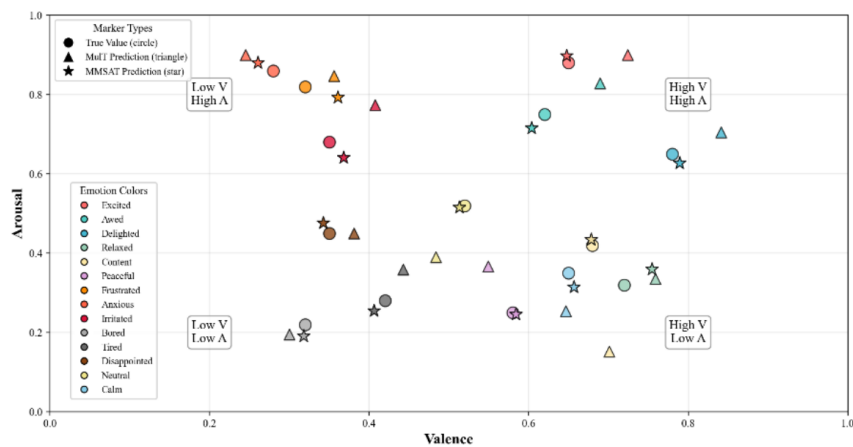


FIGURE 6. Two-dimensional spatial distribution analysis of VA space

tive, strong negative). The predicted emotion distributions from the MulT model (triangle data points) generally agree with the actual distribution; however, there are notable spatial deviations around several specific emotion points. This indicates that the MulT model is not able to as precisely identify the dimensional features of the strongest emotions. On the other hand, the predicted emotion distributions from the MMSAT model (pentagon-shaped data points) are highly overlapping with the actual distribution, showing very little variation even at the micro-level, thus indicating superior precision in its predictions. This high correlation between the predicted distributions and their actual counterparts can be attributed to the convention for multi-scale feature extraction used for the MMSAT model and the cross-modal attention mechanism it uses to extract features from visual, acoustic, and textual information. As a result, this

model accurately learns about the subtle differences between different emotions in the dimensional space and allows for accurate localization of emotions and reliable prediction of emotional states.

Model Generalization Ability and Robustness Testing

To evaluate the model's adaptability in unknown scenarios, a leave-one-out validation method is used to test the model's generalization performance in four typical rural tourism scenarios. The results are shown in Figure 7.

The cross-scenario generalization test results clearly demonstrate that the MMSAT model outperforms the MulT baseline in all test scenarios, exhibiting superior generalization ability. Specifically, in the valence dimension, MMSAT improves the CCC value by an average of 0.046 compared to MulT across all scenarios; in the arousal dimension, the

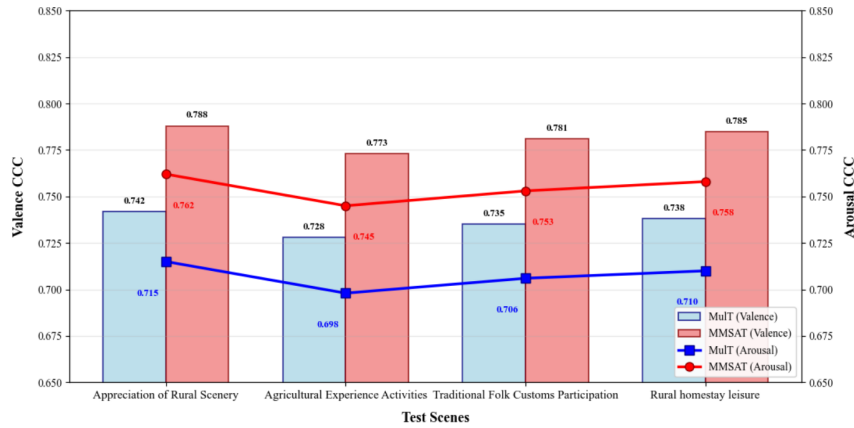


FIGURE 7. Cross-scenario generalization performance comparison

average improvement reaches 0.048. Notably, both models show relatively the lowest performance in the “farm experience activity” scenario, possibly due to the more dynamic and complex body movements involved, increasing the difficulty of emotion recognition. However, MMSAT maintains a more stable performance advantage in this challenging scenario, primarily due to its multi-scale visual feature extraction mechanism, which effectively captures multi-level information ranging from subtle facial expressions to macroscopic body movements. Furthermore, the model’s dynamic information fusion through cross-modal attention allows it to adapt to the characteristics of different scenarios, thus maintaining stable emotion recognition performance in unfamiliar environments. This result proves that the MMSAT model possesses strong scene adaptability and practical value. To evaluate the stability of the model under real-world environmental disturbances, this paper systematically tests the robustness of the model under visual noise, audio noise, and visual occlusion conditions. The results are shown in Figure 8.

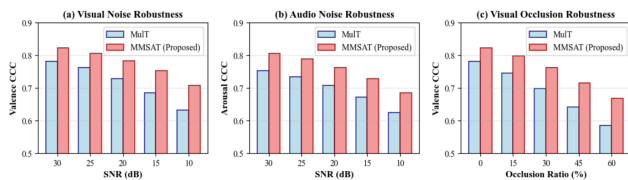


FIGURE 8. Noise and occlusion robustness test results: (a) Visual noise robustness test; (b) Audio noise robustness test; (c) Visual occlusion robustness test

In the visual noise test (Figure 8a), as the Gaussian noise intensity increases (signal-to-noise ratio decreases from 30 dB to 10 dB), the performance of both models declines, but MMSAT shows stronger anti-interference ability. Under the extreme condition of a signal-to-noise ratio of 10 dB, MMSAT’s valence CCC remains at 0.708, significantly higher than MulT’s 0.632. This is due to MMSAT’s multi-scale attention mechanism, which allows the model to automatically adjust the weight distribution of each visual scale when image quality deteriorates, reducing

reliance on features in noise-sensitive areas and instead utilizing relatively clear macroscopic scene information and mesoscopic body movement features for emotion inference. In the audio noise test (Figure 8b), MMSAT also demonstrates superior robustness. When background wind noise interference increases (signal-to-noise ratio decreases from 30 dB to 10 dB), the CCC value of MMSAT in the arousal dimension decreases from 0.805 to 0.685, a decrease of 14.9%, while MulT’s decrease reaches 16.9%. This advantage stems from the adaptive capability of the cross-modal attention mechanism: when the audio signal is severely contaminated, the model automatically reduces the weight of its dependence on the audio modality while increasing its attention to the visual and textual modalities, thereby maintaining the accuracy of emotion recognition through multimodal information complementarity. In the visual occlusion test (Figure 8c), as the occlusion ratio increases from 0% to 60%, the performance degradation of MMSAT is significantly slower than that of MulT. At a 30% occlusion ratio, MMSAT’s valence CCC is 0.762, while MulT’s has dropped to 0.698; even at a 60% occlusion ratio, MMSAT maintains a CCC value of 0.668. This indicates that the multi-scale feature extraction mechanism plays a crucial role: when certain regions of the image are occluded, the model can obtain compensatory information from other unoccluded scales. For example, when the face is partially occluded, it can still make emotion judgments based on body posture and scene context, thereby maintaining relatively stable recognition performance. The analysis in Figure 4(b) indicates a primary reliance on the visual modality, which aligns with psychological findings. However, the robustness tests (Figure 8c) reveal that performance degrades under severe visual occlusion, highlighting a practical limitation. This underscores the importance of the model’s inherent adaptive mechanism, facilitated by cross-modal attention, which dynamically redistributes attention weights towards more reliable modalities (audio, text) when visual evidence is compromised, as evidenced by the slower performance decline compared to the baseline.

V. CONCLUSION

This paper proposes a MMSAT model for rural tourism videos, addressing three major challenges in emotion recognition under complex environments through structured design, which facilitates the nuanced analysis of tourist reactions to local cultural elements. In the feature extraction stage, the model deconstructs the visual modality into macro, meso, and micro scales to capture scene atmosphere, body movements, and facial expressions, respectively; audio and text modalities are modeled with paralinguistic and semantic information using temporal CNN and BERT. In the fusion stage, an intra-modal attention mechanism achieves adaptive weighting of multi-scale visual features, while the Transformer encoder dynamically integrates heterogeneous information through cross-modal attention, ultimately outputting a continuous emotion dimension prediction. Experiments demonstrate that the model significantly outperforms existing methods in accuracy, robustness, and generalization. This research proposes a hierarchical multimodal multi-scale emotion recognition framework, validated through a dedicated rural tourism video dataset that includes diverse stimuli relevant to local cultural features. This provides a practical technical tool for optimizing rural tourism management and services, based on enhanced emotion recognition stability under complex real-world conditions.

AUTHOR CONTRIBUTIONS

Zhiguo Yang designed, collected and analyzed the data, and drafted the manuscript. Zhiguo Yang conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Zhiguo Yang participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This research was supported by, the phased research achievement of the Handan Philosophy and Social Sciences Planning Project "Research on the Digital Protection and Tourism Development of Handan Red Cultural Resources" (Project No.: XW2024263)

HUMAN RESEARCH SUBJECTS

This study was approved by the Ethics Committee of Handan University, and was performed in accordance with the principles of the Declaration of Helsinki. All eligible participants signed an informed consent form.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

- [1] L. Santamaria-Granados, J. F. Mendoza-Moreno, and G. Ramirez-Gonzalez, "Tourist recommender systems based on emotion recognition—a scientometric review," *Future Internet*, vol. 13, no. 1, pp. 2, 2020, doi: 10.3390/fi13010002.
- [2] S. Zhao, G. Jia, J. Yang, G. Ding, and K. Keutzer, "Emotion recognition from multiple modalities: Fundamentals and methodologies," *IEEE Signal Processing Magazine*, vol. 38, no. 6, pp. 59-73, 2021, doi: 10.1109/MSP.2021.3106895.
- [3] T. Zhang and Z. Tan, "Survey of deep emotion recognition in dynamic data using facial, speech and textual cues," *Multimedia Tools and Applications*, vol. 83, no. 25, pp. 66223-66262, 2024, doi: 10.1007/s11042-023-17944-9.
- [4] H. Chen, M. Wang, and Z. Zhang, "Research on rural landscape preference based on TikTok short video content and user comments," *International Journal of Environmental Research and Public Health*, vol. 19, no. 16, Art. no. 10115, 2022, doi: 10.3390/ijerph191610115.
- [5] S. K. Panda, A. K. Jena, M. R. Panda, and S. Panda, "Speech emotion recognition using multimodal feature fusion with machine learning approach," *Multimedia Tools and Applications*, vol. 82, no. 27, pp. 42763-42781, 2023, doi: 10.1007/s11042-023-15275-3.
- [6] W. Nie, Y. Yan, D. Song, and K. Wang, "Multi-modal feature fusion based on multi-layers LSTM for video emotion recognition," *Multimedia Tools and Applications*, vol. 80, no. 11, pp. 16205-16214, 2021, doi: 10.1007/s11042-020-08796-8.
- [7] M. Khan, P. N. Tran, N. T. Pham, A. El Saddik, and A. Othmani, "Mem-oCMT: Multimodal emotion recognition using cross-modal transformer-based feature fusion," *Scientific reports*, vol. 15, no. 1, pp. 5473, 2025, doi: 10.1038/s41598-025-89202-x.
- [8] L. Zhao, Y. Yang, and T. Ning, "A Three-stage multimodal emotion recognition network based on text low-rank fusion," *Multimedia Systems*, vol. 30, no. 3, pp. 142, 2024, doi: 10.1007/s00530-024-01345-5.
- [9] M. H. Yi, K. C. Kwak, and J. H. Shin, "HyFusER: Hybrid multimodal transformer for emotion recognition using dual cross modal attention," *Applied Sciences*, vol. 15, no. 3, pp. 1053, 2025, doi: 10.3390/app15031053.
- [10] B. Xie, M. Sidulova, and C. H. Park, "Robust multimodal emotion recognition from conversation with transformerbased crossmodality fusion," *Sensors*, vol. 21, no. 14, pp. 4913, 2021, doi: 10.3390/s21144913.
- [11] X. Zhang, M. Li, S. Lin, H. Xu, and G. Xiao, "Transformer-based multimodal emotional perception for dynamic facial expression recognition in the wild," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 5, pp. 3192-3203, 2023, doi: 10.1109/TCSVT.2023.3312858.
- [12] A. Chaudhari, C. Bhatt, A. Krishna, and C. M. Travieso-González, "Facial emotion recognition with inter-modalityattention-transformer-based self-supervised learning," *Electronics*, vol. 12, no. 2, pp. 288, 2023, doi: 10.3390/electronics12020288.
- [13] P. Bhattacharya, R. K. Gupta, and Y. Yang, "Exploring the contextual factors affecting multimodal emotion recognition in videos," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1547-1557, 2021, doi: 10.1109/TAFFC.2021.3071503.
- [14] P. Yang, N. Liu, X. Liu, Y. Shu, W. Ji, Z. Ren, et al., "A multimodal dataset for mixed emotion recognition," *Scientific Data*, vol. 11, no. 1, pp. 847, 2024, doi: 10.1038/s41597-024-03676-4.
- [15] W. Liu, J. L. Qiu, W. L. Zheng, and B. L. Lu, "Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 14, no. 2, pp. 715-729, 2021, doi: 10.1109/TCDS.2021.3071170.
- [16] C. Fang, H. Xiang, C. Leng, J. Chen, and Q. Yu, "Research on real-time detection of safety harness wearing of workshop personnel based on YOLOv5 and OpenPose," *Sustainability*, vol. 14, no. 10, pp. 5872, 2022, doi: 10.3390/su14105872.
- [17] H. C. Nguyen, T. H. Nguyen, J. Nowak, A. Byrski, A. Siwocha, and V. Le, "Combined YOLOv5 and HRNet for high accuracy 2D keypoint and human pose estimation," *Journal of Artificial Intelligence and Soft Computing Research*, vol. 12, no. 4, pp. 281-298, 2022, doi: 10.2478/jaiscr-2022-0019.
- [18] S. Mehra, V. Ranga, and R. Agarwal, "A deep learning approach to dysarthric utterance classification with BiLSTM-GRU, speech cue filtering, and log mel spectrograms," *The Journal of Supercomputing*, vol. 80, no. 10, pp. 14520-14547, 2024, doi: 10.1007/s11227-024-06015-x.
- [19] G. Ahmed and A. A. Lawaye, "End-to-end ASR framework for Indian-English accent: Using speech CNN-based segmentation," *International*

- Journal of Speech Technology, vol. 26, no. 4, pp. 903-918, 2023, doi: 10.1007/s10772-023-10053-w.
- [20] H. Boulal, M. Hamidi, M. Abarkan, and J. Barkani, "Amazigh CNN speech recognition system based on Mel spectrogram feature extraction method," *International Journal of Speech Technology*, vol. 27, no. 1, pp. 287-296, 2024, doi: 10.1007/s10772-024-10100-0.
 - [21] A. Dutta, G. Ashishkumar, and C. V. R. Rao, "Improving the Performance of ASR System by Building Acoustic Models using Spectro-Temporal and Phase-Based Features," *Circuits, Systems, and Signal Processing*, vol. 41, no. 3, pp. 1609-1632, 2022, doi: 10.1007/s00034-021-01848-w.
 - [22] X. Zhuang, F. Liu, J. Hou, J. Hao, and X. Cai, "Transformer-based interactive multi-modal attention network for video sentiment detection," *Neural Processing Letters*, vol. 54, no. 3, pp. 1943-1960, 2022, doi: 10.1007/s11063-021-10713-5.
 - [23] N. Pang, S. Guo, M. Yan, and C. A. Chan, "A short video classification framework based on cross-modal fusion," *Sensors*, vol. 23, no. 20, pp. 8425, 2023, doi: 10.3390/s23208425.
 - [24] L. H. Lee, J. H. Li, and L. C. Yu, "Chinese EmoBank: Building valence-arousal resources for dimensional sentiment analysis," *Transactions on Asian and Low-Resource Language Information Processing*, vol. 21, no. 4, pp. 1-18, 2022, doi: 10.1145/3489141.
 - [25] M. Yik, C. Mues, I. N. L. Sze, P. Kuppens, F. Tuerlinckx, K. De Roover, et al., "On the relationship between valence and arousal in samples across the globe," *Emotion*, vol. 23, no. 2, pp. 332, 2023, doi: 10.1037/emo0001095.
 - [26] G. S. Kumar, N. Sampathila, and R. J. Martis, "Classification of human emotional states based on valence-arousal scale using electroencephalogram," *Journal of Medical Signals & Sensors*, vol. 13, no. 2, pp. 173-182, 2023, doi: 10.4103/jmss.jmss_169_21.
 - [27] R. G. Praveen, P. Cardinal, and E. Granger, "Audio-visual fusion for emotion recognition in the valence-arousal space using joint cross-attention," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 5, no. 3, pp. 360-373, 2023, doi: 10.1109/TBIOM.2022.3233083.
 - [28] M. Awatef, B. Hayet, and L. Zied, "Multimodal emotion recognition: Integrating speech and text for improved valence, arousal, and dominance prediction," *Annals of Telecommunications*, vol. 80, no. 5, pp. 401-415, 2025, doi: 10.1007/s12243-025-01069-1.
 - [29] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, et al., "Dawn of the transformer era in speech emotion recognition: Closing the valence gap," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10745-10759, 2023, doi: 10.1109/TPAMI.2023.3263585.
 - [30] C. Li, L. Xie, X. Wang, H. Pan, and Z. Wang, "A disentanglement mamba network with a temporally slack reconstruction mechanism for multimodal continuous emotion recognition," *Multimedia Systems*, vol. 31, no. 2, pp. 1-18, 2025, doi: 10.1007/s00530-025-01758-w.
 - [31] H. Lian, C. Lu, S. Li, Y. Zhao, C. Tang, and Y. Zong, "A survey of deep learning-based multimodal emotion recognition: Speech, text, and face," *Entropy*, vol. 25, no. 10, pp. 1440, 2023, doi: 10.3390/e25101440.
 - [32] X. Meng, "Cross-domain information fusion and personalized recommendation in artificial intelligence recommendation system based on mathematical matrix decomposition," *Scientific Reports*, vol. 14, no. 1, pp. 7816, 2024, doi: 10.1038/s41598-024-57240-6.
 - [33] W. Ai, Y. Shou, T. Meng, and K. Li, "Der-gcn: Dialog and event relation-aware graph convolutional neural network for multimodal dialog emotion recognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 3, pp. 4908-4921, 2024, doi: 10.1109/TNNLS.2024.3367940.
 - [34] L. Guo, L. Wang, J. Dang, Y. Fu, J. Liu, and S. Ding, "Emotion recognition with multimodal transformer fusion framework based on acoustic and lexical information," *IEEE MultiMedia*, vol. 29, no. 2, pp. 94-103, 2022, doi: 10.1109/MMUL.2022.3161411.
 - [35] Y. Wu, M. Daoudi, and A. Amad, "Transformer-based self-supervised multimodal representation learning for wearable emotion recognition," *IEEE Transactions on Affective Computing*, vol. 15, no. 1, pp. 157-172, 2023, doi: 10.1109/TAFFC.2023.3263907.