MS^{2}FL: Modality-shared and Modality-specific Feature Learning for Multimodal Emotion Recognition

XINHUI LI, Anhui Province Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University, China

HAO CHEN, Anhui Province Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University, China

MINCHAO WU, School of Computer Science and Artificial Intelligence, Hefei Normal University, China

QINGWEI SONG, Anhui Province Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University, China

FAN LI, CAAC Key Laboratory of Civil Aviation Flight Technology and Flight Safety, Civil Aviation Flight University of China, China

JINYANG HUANG, School of Computer Science and Information Engineering, Hefei University of Technology, China

FENGQI CUI, Institute of Advanced Technology, University of Science and Technology of China, China

ZHAO LV*, Anhui Province Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University, China

Multimodal emotion recognition based on complementary physiological signals such as electroencephalogram (EEG) and eye movements can effectively reflect human emotional states, demonstrating significant potential in fields such as rehabilitation monitoring and driving safety. However, existing multimodal emotion recognition approaches often focus solely on either feature fusion or feature alignment strategies, lacking a unified framework capturing both complementary and modality-specific representations. To address this limitation, we propose a modality-shared and modality-specific feature learning framework (MS^{2}FL) for multimodal emotion recognition. Specifically, a dual-path encoder is employed first to extract the spatiotemporal frequency features of EEG signals and dynamic response features of eye movement signals, respectively. Then, a specificity-preserving feature distribution alignment

*Corresponding author.

Authors' Contact Information: Xinhui Li, xinhuili@ahu.edu.cn, Anhui Province Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University, Hefei, Anhui, China; Hao Chen, e23201026@stu.ahu.edu.cn, Anhui Province Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University, Hefei, Anhui, China; Minchao Wu, wu$_m$in$_c$hao@hotmail.com, School of Computer Science and Artificial Intelligence, Hefei Normal University, Hefei, Anhui, China; Qingwei Song, e23201118@stu.ahu.edu.cn, Anhui Province Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University, Hefei, Anhui, China; Fan Li, lifan@cafuc.edu.cn, CAAC Key Laboratory of Civil Aviation Flight Technology and Flight Safety, Civil Aviation Flight University of China, Guanghan, Sichuan, China; Jinyang Huang, hjy@hfut.edu.cn, School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, China; FengQi Cui, fengqi$_c$ui@mail.ustc.edu.cn, Institute of Advanced Technology, University of Science and Technology of China, Hefei, China; Zhao Lv, emsp$_a$hu@163.com, Anhui Province Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University, Hefei, Anhui, China.

Manuscript submitted to ACM

mechanism is introduced to alleviate distribution discrepancies between modalities while reinforcing modality-specific modeling of distinctive features. Finally, a feature distribution enhancement fusion strategy is utilized to effectively integrate the shared features across modalities, thus improving the representational capacity of the fused features. Experimental results demonstrate that MS^{2}FL achieves recognition accuracies of 96.99% and 89.83% on the SEED and SEED-V datasets, respectively, significantly improving emotion recognition performance. This study provides a concise and effective solution for emotion perception neurotechnology and offers a practical solution for emotion perception health monitoring and rehabilitation.

CCS Concepts: • **Computing methodologies** → **Emotion recognition**; *Multimodal learning*; Feature selection; Neural networks.

Additional Key Words and Phrases: Multimodal Emotion Recognition, Brain–Computer Interface, Electroencephalogram, Eye Movements

ACM Reference Format:

Xinhui Li, Hao Chen, Minchao Wu, Qingwei Song, Fan Li, Jinyang Huang, FengQi Cui, and Zhao Lv. 2026. MS^{2}FL: Modality-shared and Modality-specific Feature Learning for Multimodal Emotion Recognition. *J. ACM* 37, 4, Article 111 (August 2026), 25 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Emotion plays a crucial role in regulating human behavior [1], and its automatic recognition is of great significance in a variety of application scenarios, including natural human-computer interaction [2], mental state monitoring and assisted diagnosis [3, 4], and driving safety management [5]. Existing methods can be broadly classified into physiological signal-based and non-physiological data-based recognition approaches. The former encompasses modalities such as electroencephalogram (EEG) [6], eye movements [7], electromyography (EMG)[8], and electrocardiography (ECG)[9], whereas the latter includes video [10], speech [11], and text [12]. Compared with non-physiological data, physiological signals provide a more objective reflection of an individual's internal emotional state and offer high reliability due to their resistance to manipulation. Among these, EEG, which directly reflects neural brain activity and is closely linked to fundamental neural mechanisms [13], has become one of the most extensively studied signals in emotion recognition tasks [14]. Nevertheless, due to the inherent complexity of emotional states, relying solely on a single modality often falls short of achieving accurate detection. EEG signals not only provide stable neural patterns but also enable fine-grained characterization of subtle emotional fluctuations, owing to their high temporal resolution [15]. In addition, eye movements can capture behavioral features associated with environmental perception and cognitive processing [16]. Owing to their complementary characteristics, combining EEG and eye movements has attracted growing interest in recent studies [17], resulting in improved emotion discrimination and broader application prospects. Benchmark datasets commonly used to evaluate multimodal emotion recognition include SEED [18], which contains EEG and eye movement data from 12 subjects corresponding to three emotion categories: Sad, Neutral, and Happy, and its extended version SEED-V [19], which includes 16 subjects and covers five emotion categories: Disgust, Fear, Sad, Neutral, and Happy. These datasets provide a standard basis for assessing the effectiveness of multimodal approaches, motivating the adoption of advanced deep learning techniques for feature extraction and representation.

Deep learning models have been widely adopted in multimodal emotion recognition research due to their powerful capabilities in feature extraction and representation learning [20]. Existing methods based on deep learning can be broadly categorized into two types. One class involves feature fusion strategies, where features from different modalities are typically concatenated or weighted and then fed into a shared sub-network for high-level abstraction and emotion classification [21]. For instance, Liu et al. [22] proposed a bimodal deep autoencoder (BDAE) to learn shared representations from EEG and eye-tracking data to improve recognition performance. Zhang et al. [23] integrated

heterogeneous convolutional neural networks (HCNNs) with multimodal factorized bilinear pooling (MFB) to extract and fuse modality-specific features, followed by ensemble optimization. However, due to distribution differences among features of different modalities, such methods often overlook this heterogeneity, and directly fusing unaligned features can lead to representation conflicts, which makes it difficult to exploit complementary information effectively. The second category emphasizes inter-modal alignment mechanisms, aiming to explicitly match features through attention or gating structures. For example, Chen et al. [24] implemented word-level alignment using gated LSTM and temporal attention, while Xu et al. [25] built temporal correspondences between modalities based on temporal attention. However, these methods often mix unimodal features with cross-modal relational information during the alignment process, neglecting the independent expressive ability of the original modality in emotion recognition, which tends to cause the loss of modality-specific information and thus weakens the overall model's ability to represent complex emotional states and its generalization performance. Therefore, effectively combining feature fusion and alignment strategies to simultaneously model modality complementarity and modality specificity remains a challenge.

To address the above issues, this paper presents a novel modality-shared and modality-specific feature learning (MS²FL) framework, which aims to simultaneously model modality complementarity and modality specificity. Following an "alignment first, then fusion" strategy, MS²FL alleviates distributional heterogeneity across modalities and enables more effective integration of emotional information. In addition to independent feature extractors for EEG and eye movement signals, the framework mainly consists of two key modules. The first is the specificity-preserving feature distribution alignment (SP-FDA) module, which maps features into a shared latent space and introduces a contrastive and distribution consistency loss (EEGEye-CLDC Loss). This module constrains constrains the distribution consistency between modalities through contrastive learning while enhancing the modality-specific modeling of unimodal features. The second is the feature distribution enhancement fusion (FDEF) module, which enhances the peak-valley distribution differences of common features in the semantic space to improve discriminability, and introduces a multi-modal distillation to single-modal classification loss (MD2SC Loss). This module achieves effective fusion that accounts for modality complementarity and independence by concatenating the enhanced features. In summary, MS²FL not only preserves the independence of single modalities but also mines the complementarity between modalities, thereby improving emotion recognition accuracy and model generalization.

The main contributions of this paper are as follows:

- We propose MS²FL, which effectively integrates feature fusion and alignment strategies, enabling simultaneous modeling of modality complementarity and modality specificity.
- We design the SP-FDA module to align inter-modality feature distributions while enhancing the modality-specific modeling of unimodal features.
- We design the FDEF module to achieve effective fusion that captures modality complementarity and modality specificity.
- Extensive experiments conducted on the SEED and SEED-V datasets validate that MS²FL demonstrates superior performance compared to existing state-of-the-art (SOTA) approaches in multimodal emotion recognition tasks.

2 Related Works

2.1 EEG-based Emotion Recognition

EEG has become a widely adopted modality in emotion recognition research due to its high temporal resolution, non-invasiveness, rapid acquisition, and low implementation cost, contributing to the development of numerous EEG-based

methods [14]. Due to the weak amplitude and highly non-stationary nature of EEG signals, early research typically relied on manual extraction of time-frequency features such as differential entropy (DE) and power spectral density (PSD) to improve recognition stability [26]. These handcrafted features laid the foundation for subsequent advances in EEG-based emotion recognition.

With the rise of deep learning, convolutional neural networks (CNNs), known for their strong local feature extraction capabilities, have been widely applied to directly model DE and PSD features, enabling automatic extraction of local spatiotemporal patterns and enhancing emotional representation. For example, Li *et al.*[27] organized EEG differential entropy features into three-dimensional maps based on electrode spatial distribution and used CNNs to extract frequency-spatial features. Chao *et al.*[28] further incorporated frequency, spatial, and band-specific information to construct multi-band feature matrices and applied capsule networks for emotion classification. In addition, due to the irregular spatial configuration of EEG electrodes, recent studies have explored graph convolutional networks (GCNs) to better capture spatial dependencies in unstructured EEG data [29]. These models represent EEG channels as graph nodes, with edges encoding inter-channel relationships [30]. For instance, Song *et al.*[31] utilized dynamical GCNs (DGCNNs) to dynamically model intrinsic dependencies among EEG channels. Li *et al.*[32] proposed CAGLE-net, which leverages correlation analysis to guide graph construction and integrates locally enhanced features, achieving state-of-the-art performance on the SEED dataset.

In summary, EEG-based emotion recognition has made significant strides by leveraging both handcrafted features and deep learning techniques, leading to more effective modeling of spatiotemporal dynamics and inter-channel relationships. As EEG relationship modeling continues to show great potential, recent studies have further explored integrating complementary physiological modalities. Due to the limitations of unimodal methods in comprehensively capturing the multidimensional characteristics and complex variations of emotions, multimodal fusion approaches have been proposed to improve the accuracy and robustness of emotion recognition by integrating diverse physiological signals [33].

2.2 Multimodal Emotion Recognition

To address the inherent limitations of single-modal approaches, multimodal emotion recognition has gained increasing attention for its ability to capture more comprehensive and nuanced emotional information. By combining signals from different physiological or behavioral sources, multimodal systems can exploit the complementary characteristics of each modality. Among them, feature fusion and feature alignment are two core strategies in multimodal emotion recognition. Feature fusion focuses on effectively combining features from different modalities to obtain more discriminative emotional representations. In contrast, feature alignment aims to establish a consistent feature space across modalities to reduce distributional discrepancies among heterogeneous data.

2.2.1 Feature Fusion. Feature fusion strategies are generally categorized into direct fusion and adaptive fusion. Among direct fusion methods, concatenation is the most commonly used technique, where features from different modalities are concatenated into a high-dimensional feature vector and then input into a classifier for emotion recognition. Although this method is simple to implement, it may not fully exploit inter-modal interactions, which can limit the model performance [34]. Another approach is weighted fusion, which assigns weights to features from different modalities and computes a weighted sum to emphasize the contribution of critical modalities. For example, Yan et al. [35] proposed a weighted fusion strategy that assigns three types of weights to different signal modalities by analyzing intra- and inter-modal information, thereby improving emotion classification performance.

Compared to direct fusion, adaptive fusion approaches dynamically adjust feature weights during the fusion process through deep learning models or attention mechanisms to enhance fusion effectiveness. Attention-based fusion enables the model to automatically allocate inter-modal weights, thereby focusing on more discriminative features. For instance, Liu et al. [21] proposed an attention-based fusion strategy within the deep canonical correlation analysis (DCCA) for multimodal emotion recognition. Li et al. [36] designed a unified cross-attention module named token-channel compound (TACO) to capture inter-modal dependencies at both the channel and token levels. Guo et al. [37] developed a framework that integrates sparse self-attention, channel attention, and spatial attention to enhance the semantic information across feature map layers, thereby improving the expressive capacity of fused features. In addition, deep model-based fusion methods employ deep learning architectures to facilitate feature interactions, allowing features from different modalities to be fully integrated. For example, Tang et al. [38] proposed a hierarchical multimodal network (RHPRNet) that leverages deep architectures to perform spatial-frequency feature extraction, cross-modal encoding, and hierarchical fusion, enhancing the accuracy of multimodal emotion recognition. Li et al. [39] utilized a cross-modal transformer (CMT) to optimize auxiliary modality features, adopted low-rank fusion to reduce redundancy, and enhanced primary modality features via an improved CMT, leveraging self-attention to exploit inter-modal commonalities and complementarities.

2.2.2 Feature Alignment. Unlike feature fusion, feature alignment methods primarily address the heterogeneity of multimodal data by mapping features from different modalities into a shared feature space, thus minimizing inter-modal distribution gaps. In multimodal emotion recognition tasks, alignment is typically achieved through two main strategies: forced alignment at the time-step level and sequence dependency-based alignment. Forced time-step alignment refers to dividing signals from different modalities into synchronized time units—often aligned with word or frame boundaries—upon which inter-modal interaction and fusion are performed. This method emphasizes local temporal correspondence across modalities. For instance, Gu et al.[40] proposed a hierarchical attention strategy for word-level alignment to improve emotional inference. Rohanian et al.[41] introduced a time-dependent recurrent approach that progressively fuses multimodal features at each word, enabling dynamic integration over time. Huddar et al. [42] presented an emotion recognition model that extracts features at the word level and performs inter-modal interaction based on a forced alignment mechanism.

In contrast, sequence dependency-based alignment methods emphasize modeling global relationships and structural consistency across modalities, leveraging complementary information to enhance alignment effectiveness. For example, Chen et al. [43] introduced a Time and semantic interaction network (TSIN), which models temporal correlations among multimodal signals using a fine-grained temporal alignment mechanism to enhance cross-modal temporal consistency. Shou et al. [44] proposed the alignment and graph fusion via information bottleneck (AGF-IB) method, which addresses modality heterogeneity via adversarial alignment and graph contrastive learning, and optimizes cross-modal information exchange based on the information bottleneck theory. Meng et al. [45] proposed the masked graph learning with recursive alignment (MGLRA) method, which employs a recursive alignment module to progressively eliminate inter-modal discrepancies and introduces a masked GCN for feature fusion, thereby improving the alignment process and overall model performance in multimodal emotion recognition. Similarly, Wang et al. [46] proposed a multi-modal domain adaptive variational autoencoder (MMDA-VAE), which leverages adversarial learning and cycle-consistency regularization to learn a shared cross-domain latent representation between EEG and eye movement data for feature alignment.

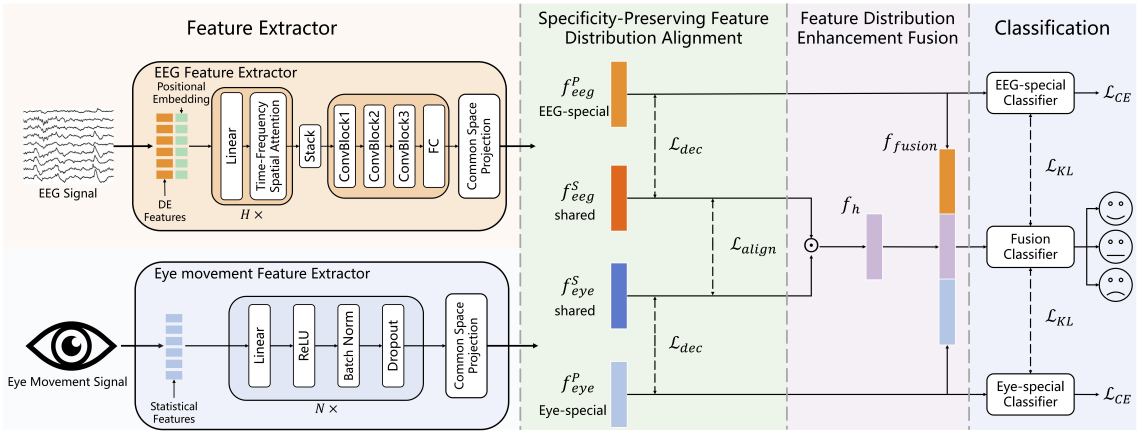

Fig. 1. The proposed modality-shared and modality-specific feature learning multimodal fusion framework. H represents the number of heads in the multi-head attention, N represents the number of layers in the eye movement signal feature extractor, and $\odot$ represents the Hadamard product.

While feature fusion methods can effectively integrate multimodal information, their lack of explicit alignment mechanisms leads to degraded fusion performance due to ignoring inter-modal distribution differences that cause representation conflicts. In contrast, feature alignment approaches can reduce distributional discrepancies among modalities but may weaken the independent representational capacity of unimodal features, leading to potential information loss [47]. Therefore, how to combine feature fusion and feature alignment methods to jointly model modality complementarity and modality specificity, thereby further improving the performance of multimodal emotion recognition, remains a promising yet challenging research direction.

3 Method

This paper presents a modality-shared and modality-specific feature learning framework for multimodal emotion recognition to improve the complementary modeling and independent representation of EEG and eye movement signals in emotion recognition. As shown in Fig. 1, the framework consists of three main components: feature extractors, SP-FDA, and FDEF. The dual-branch feature extractor captures the spatiotemporal frequency features of EEG and the dynamic response features of eye movements, enabling deep modeling of modality-specific information. SP-FDA maps EEG and eye features into a unified semantic space, introduces the EEGEye-CLDC Loss to ensure distribution consistency of shared features, and applies contrastive learning between shared and specific features to enhance alignment and preserve modality characteristics. In the fusion stage, FDEF strengthens peak-valley distribution differences to improve discriminability, while MD2SC Loss, applied to concatenated specific features, enhances the perception of category structure and generalization. Module details are elaborated in the following sections.

3.1 Feature Extractor

3.1.1 EEG Features. In emotion recognition, temporal, frequency, and spatial features characterize brain neural activity under emotional states from different perspectives, and are essential for comprehensively modeling emotional representation patterns. Temporal features reflect the dynamic variations of EEG signals, frequency features reveal the

distribution of emotion-related neural rhythms, and spatial features capture the co-activation and spatial dependencies among different brain regions. Effectively integrating these three types of features facilitates comprehensive modeling of the brain's complex response patterns during emotional states.

Previous studies have demonstrated that DE has been widely investigated and utilized due to its effectiveness in extracting emotional representations from EEG signals [48]. Inspired by this, we adopt DE as the fundamental feature representation for the EEG modality to enhance the expression of emotional information in EEG signals. Assuming that EEG signals follow a Gaussian distribution, the DE feature is computed as follows:

$$DE = \frac{1}{2} \ln \left(2\pi e \sigma^2\right), \tag{1}$$

where σ^2 denotes the signal variance within a specific time window. This paper employs a 4-second sliding window to extract DE features, which are partitioned into five canonical frequency bands: delta (1–4 Hz), theta (4–8 Hz), alpha (8–14 Hz), beta (14–31 Hz), and gamma (31–50 Hz), to sufficiently capture emotion-related multi-band neural activity patterns. The multi-band DE feature X_{de} can be expressed as:

$$X_{de} = [DE_\delta, DE_\theta, DE_\alpha, DE_\beta, DE_\gamma], \tag{2}$$

where DE_δ to DE_γ represent the differential entropy features in the five frequency bands respectively. During model training, the extracted DE features $X_{de} \in \mathbb{R}^{B\times C\times 5}$ contain three dimensions: batch size (B), EEG channels (C), and frequency bands (5).

To further integrate the aforementioned multidimensional features and enhance spatial modeling capabilities, this paper proposes an EEG feature extractor based on temporal-frequency-spatial attention modules, inspired by the small-world network theory in neuroscience [49] and graph attention networks (GAT) [50]. This method learns joint temporal-frequency-spatial representations among electrodes in an end-to-end manner, mapping 1D DE features with positional encoding into 2D grid-like structures. The approach preserves the temporal-frequency dynamics of raw signals while modeling nonlinear coupling relationships between electrode channels, adaptively capturing multi-electrode coordinated response patterns in emotional encoding, thereby providing highly discriminative inputs for subsequent convolutional networks. The complete architecture consists of two components: a temporal-frequency-spatial attention module and a convolutional module.

The temporal-frequency-spatial attention module initially receives EEG signals and their corresponding positional information. The positional encoding $E \in \mathbb{R}^{C\times 3}$ is derived from 3D electrode coordinates in the international 10–20 system. To ensure effective integration of EEG signals with spatial information at each timestep, the positional encoding E is expanded along the batch dimension to match the EEG signal dimensionality, resulting in $E_{\text{batch}} \in \mathbb{R}^{B\times C\times 3}$. This expanded encoding is then concatenated with X_{de} to form an enriched feature representation $X_{\text{embed}} \in \mathbb{R}^{B\times C\times 8}$.

$$X_{\text{embed}} = \text{concat}(X_{de}, E_{\text{batch}}), \tag{3}$$

where X_{de} denotes the DE features of EEG signals and E_{batch} represents the electrode channel positional encoding of EEG signals.

The concatenated features undergo linear projection to match multi-head attention requirements. This projected representation is then nonlinearly transformed through a ReLU activation layer to enhance feature discriminability:

$$F_{\text{embed}} = \text{ReLU}(W \cdot X_{\text{embed}}), \tag{4}$$

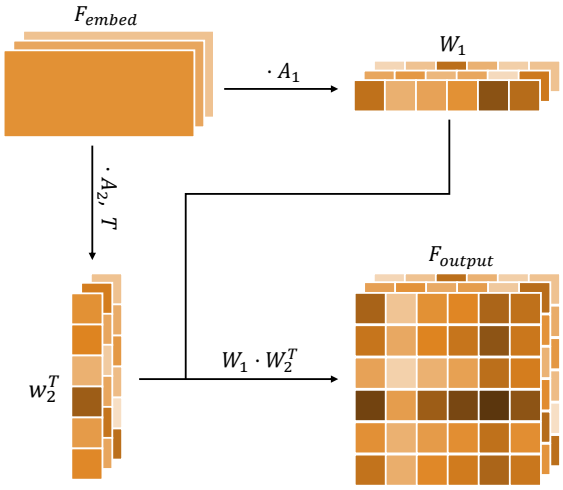

Fig. 2. Temporal-frequency-spatial attention mechanism. Multi-head attention captures time–frequency and channel dependencies, and aggregates the outputs to generate unified feature representations.

where $W \in \mathbb{R}^{8\times(H\times D_1)}$ denotes the weight matrix of the linear transformation, H specifies the number of attention heads, and D_1 represents the embedding dimension per head. After mapping and activation, a reshape operation yields $F_{\text{embed}} \in \mathbb{R}^{B\times H\times C\times D_1}$ as input for subsequent attention computation.

The module employs multi-head attention mechanism to capture multi-level relationships within input features. As shown in Fig. 2, the transformed features establish relationships from multiple perspectives through parallel computation heads.

For each attention head, two matrices $W_1, W_2 \in \mathbb{R}^{B\times C\times 1}$ are generated via matrix multiplication. Their interaction computes a relation matrix $e \in \mathbb{R}^{B\times C\times C}$, which characterizes attention weights between electrode channels. The final output $F_{\text{output}} \in \mathbb{R}^{B\times H\times C\times C}$ is obtained by stacking relation matrices from all heads:

$$W_1 = F_{\text{embed}} \cdot A_1, \quad W_2 = F_{\text{embed}} \cdot A_2, \tag{5}$$

$$e = W_1 \cdot W_2^{\top}, \tag{6}$$

$$F_{\text{output}} = \text{Stack}(e_1, e_2, \ldots, e_H), \tag{7}$$

where $A_1, A_2 \in \mathbb{R}^{B\times D_1\times 1}$ are trainable parameter matrices, and $\{e_1, e_2, \ldots, e_H\}$ denote relation matrices computed by each attention head.

Due to the relatively small size of the attention matrices, using deeper networks (e.g., ResNet50) may increase computational complexity and even lead to overfitting. Therefore, the convolutional module employs a lightweight convolutional neural network to extract deep features from the generated attention matrices. Through this design, it not only avoids the computational burden caused by deep networks but also effectively extracts deep feature information.

3.1.2 Eye Movement Features. In Emotion Recognition research, heterogeneous data modalities require tailored feature extraction approaches. For low-dimensional eye movement signals, we design a dedicated MLP architecture to preserve raw discriminative features while extracting deep representations, thereby providing robust support for emotion analysis. Following [51], the input consists of 33-dimensional eye movement features encompassing key physiological indicators like pupil diameter and divergence. The MLP interleaves batch normalization and Dropout layers between linear transformations to accelerate convergence and prevent overfitting, effectively mining emotion-relevant patterns from eye movement data.

3.2 Specificity-Preserving Feature Distribution Alignment

EEG and eye movement data stem from fundamentally different sources, leading to discrepancies in their feature distributions. Moreover, each modality independently provides valuable emotional cues, but existing alignment methods often overemphasize cross-modal consistency, neglecting the unique expressive power of single modalities. To address this, we propose SP-FDA, which maps shared features into a unified space. Specifically, we introduce EEGEye-CLDC Loss, which consists of two main components: one that preserves modality-specific discriminative characteristics, and another that constrains the distribution consistency of modality-shared features.

First, modality-specific features are extracted from EEG and eye movement signals through their respective feature extractors, denoted as $f^P_{\text{eeg}}, f^P_{\text{eye}} \in \mathbb{R}^{B\times D_2}$. Subsequently, each modality feature is projected into a unified semantic space via independent shared-space projection modules, resulting in the corresponding shared feature representations $f^S_{\text{eeg}}, f^S_{\text{eye}} \in \mathbb{R}^{B\times D_2}$. Here, D_2 denotes the embedding dimension of the extracted features. This mapping process is designed to minimize semantic expression discrepancies across modalities, thereby laying a foundation for feature distribution alignment and enhancing the modeling capability of complementary information during the fusion stage.

After completing modality-specific feature extraction and shared space projection, both EEG and eye movement signals obtain their corresponding specific and shared feature representations. To achieve cross-modal feature alignment and unimodal-specific modeling simultaneously, this paper further introduces the EEGEye-CLDC Loss, which jointly constrains inter-modality consistency and intra-modality independence. Specifically, the specific features are directly derived from the original modality data, preserving fine-grained modality-specific patterns with strong discriminative capability. In contrast, the shared features are obtained through projection into the common space and primarily capture shared semantic information across modalities in the emotional dimension, serving as the core foundation for multimodal fusion.

The EEGEye-CLDC Loss consists of three components: inter-modality alignment loss (including contrastive loss and MMD-based distribution consistency constraint) and intra-modality decoupling loss. First, to enhance the discriminability and cross-modal matching ability of the shared features, the inter-modality alignment loss incorporates a bidirectional contrastive loss, which reinforces the shared features' semantic consistency and alignment across modalities through contrastive learning. Given the shared features f^S_{eeg} and f^S_{eye}, the contrastive loss between them is defined as follows:

$$\mathcal{L}_{\text{con}} = -\frac{1}{2B}\sum_{i=1}^{B}\left[\log\frac{\exp\left(f^S_{\text{eeg},i}\cdot f^S_{\text{eye},i}/\tau\right)}{\sum_{j=1}^{B}\exp\left(f^S_{\text{eeg},i}\cdot f^S_{\text{eye},j}/\tau\right)} + \log\frac{\exp\left(f^S_{\text{eye},i}\cdot f^S_{\text{eeg},i}/\tau\right)}{\sum_{j=1}^{B}\exp\left(f^S_{\text{eye},i}\cdot f^S_{\text{eeg},j}/\tau\right)}\right], \tag{8}$$

where τ denotes the temperature parameter, I represents the identity label corresponding to each sample index.

To further reduce distribution discrepancy in inter-modal shared features, we introduce the Maximum Mean Discrepancy (MMD) loss:

$$\begin{aligned}\mathcal{L}_{\text{mmd}} =& \mathbb{E}\left[k(f_{\text{eeg}}^S, f_{\text{eeg}}^S)\right] + \mathbb{E}\left[k(f_{\text{eye}}^S, f_{\text{eye}}^S)\right] \\ &- 2\mathbb{E}\left[k(f_{\text{eeg}}^S, f_{\text{eye}}^S)\right],\end{aligned} \tag{9}$$

where $k(\cdot,\cdot)$ is the kernel function for measuring similarity between samples, and the multi-scale RBF kernel [52] is employed. The final modality alignment loss is jointly constrained by the bidirectional contrastive learning loss and MMD:

$$\mathcal{L}_{\text{align}} = \mathcal{L}_{\text{con}} + \lambda_{\text{mmd}} \cdot \mathcal{L}_{\text{mmd}}. \tag{10}$$

To ensure the independence between shared features and specific features, a decoupling loss term is introduced by penalizing the Euclidean distance between them, thereby enhancing the independent representation capability of specific features:

$$\begin{aligned}\mathcal{L}_{\text{dec}} =& \mathbb{E}\left[\exp\left(-\left\|f_{\text{eeg}}^S - f_{\text{eeg}}^P\right\|_2\right)\right] \\ &+ \mathbb{E}\left[\exp\left(-\left\|f_{\text{eye}}^S - f_{\text{eye}}^P\right\|_2\right)\right].\end{aligned} \tag{11}$$

The EEGEye-CLDC Loss is formulated as:

$$\begin{aligned}\mathcal{L}_{\text{EEGEye-CLDC}} &= \mathcal{L}_{\text{align}} + \lambda_{\text{dec}} \cdot \mathcal{L}_{\text{dec}} \\ &= \mathcal{L}_{\text{con}} + \lambda_{\text{mmd}} \cdot \mathcal{L}_{\text{mmd}} + \lambda_{\text{dec}} \cdot \mathcal{L}_{\text{dec}},\end{aligned} \tag{12}$$

where λ_{mmd} and λ_{dec} are weight hyperparameters for MMD and decoupling losses respectively, controlling each component's contribution.

Through the EEGEye-CLDC Loss, SP-FDA jointly enhances cross-modal alignment and modality specificity. The contrastive term $\mathcal{L}_{\text{con}}$ enforces semantic consistency between shared features across modalities, strengthening their mutual alignment, while the MMD term $\mathcal{L}_{\text{mmd}}$ further mitigates distribution discrepancies among the shared features, ensuring a consistent semantic space. Concurrently, the decoupling term $\mathcal{L}_{\text{dec}}$ preserves fine-grained modality-specific patterns by penalizing overlap with shared features, maintaining specificity representation capability. Collectively, these components establish a stable, robust, and discriminative feature foundation, substantially improving the performance and generalization of subsequent multimodal fusion for emotion recognition.

3.3 Feature Distribution Enhancement Fusion Module

After acquiring both shared and specific features from EEG and eye movement modalities, effective fusion emerges as a critical challenge. Conventional fusion strategies typically adopt simple concatenation or weighted mechanisms, which fail to simultaneously model inter-modal complementarity and intra-modal independence, often resulting in information dilution of specific features and constrained model discriminability. To address this limitation, we propose FDEF designed to better coordinate shared and specific features across modalities, thereby enhancing the discriminability and representational capacity of fused features.

The framework achieves this through two principal mechanisms: enhancing the peak-valley distribution discrepancy of shared features in semantic space to amplify their discriminability, and concatenating modality-specific features while introducing the MD2SC Loss that transfers multimodal information to single-modality specific features via distillation strategy, coupled with single-modality classification supervision. This dual approach strengthens the discriminative capability and generalization performance of individual modalities, ultimately achieving high-efficiency fusion of inter-modal complementarity and intra-modal independence.

EEG and eye movement signals are initially processed through feature extractors and a shared space projection module to obtain common feature representations in a unified dimensionality. To emphasize the peak-valley characteristics of the feature distributions and to better exploit inter-modal complementarity, we employ the Hadamard product for element-wise multiplication between the features of the two modalities. This multiplicative interaction differs from additive operations by amplifying larger feature values while reducing smaller ones in relative magnitude, thereby expanding the dynamic range of the fused representations. As a result, distinctive patterns across modalities are accentuated, which enhances the discriminability of the shared features and facilitates improved classification performance. This operation can be formulated as:

$$f_h = f_{\text{eeg}}^S \odot f_{\text{eye}}^S, \tag{13}$$

where f_{eeg}^S and f_{eye}^S denote the shared features of EEG and eye movement signals respectively, $\odot$ represents the element-wise multiplication operation, and $f_h \in \mathbb{R}^{B\times D_2}$ indicates the enhanced shared features after preliminary fusion.

Although the Hadamard product operation effectively enhances the discriminative capacity of multimodal shared features, relying solely on fused representations may reduce the expressive power of single-modality features and potentially compromise generalization. To address this, we adopt a residual-style fusion strategy to explicitly preserve unimodal information while stabilizing the overall feature representation. Specifically, we concatenate the enhanced shared features f_h obtained from the Hadamard product with EEG-specific features f_{eeg}^P and eye-movement-specific features f_{eye}^P, constructing the final fused feature vector $f_{\text{fusion}} \in \mathbb{R}^{B\times 3D_2}$. This design allows the model to retain unique modality-specific characteristics while simultaneously benefiting from enhanced inter-modal interactions, improving both discriminability and stability. The concatenation process is formulated as:

$$f_{\text{fusion}} = \text{concat}(f_{\text{eeg}}^P, f_h, f_{\text{eye}}^P). \tag{14}$$

The fused features are then mapped through a linear layer, where the Softmax function processes the features to produce classification results:

$$\hat{y} = \text{Softmax}(W_{\text{cls}} f_{\text{fusion}} + b_{\text{cls}}), \tag{15}$$

where W_{cls} and b_{cls} denote the classifier's weight matrix and bias term respectively, with $\hat{y}$ representing the predicted class probabilities.

To further enhance the discriminative capacity and generalization capability of fused features, we introduce the MD2SC Loss. This loss function primarily strengthens the discriminative power of single-modality specific features, thereby improving fused feature performance. Specifically, the MD2SC Loss integrates weak distillation with strong classification supervision to emphasize the importance of single-modality specific features in multimodal learning, while also enhancing the ability of feature extractors to mine information from individual modalities.

Inspired by the concept of knowledge distillation, we treat the class prediction $\hat{y}$ of the fused feature f_{fusion} as a soft label to guide the prediction results $\hat{y}_{\text{eeg}}$ and $\hat{y}_{\text{eye}}$ of the modality-specific features f_{eeg}^{P} and f_{eye}^{P} toward alignment. his forces individual modalities to learn globally consistent discriminative information during training. The process achieves "weak distillation" from the multimodal representation to each single modality by minimizing the Kullback–Leibler divergence between the fused prediction distribution and those of individual modalities, effectively enhancing the representational capacity and discriminability of the modality-specific features. The distillation loss is defined as:

$$\mathcal{L}_{\text{KL}} = D_{\text{KL}}(\hat{y} \parallel \hat{y}_{\text{eeg}}) + D_{\text{KL}}(\hat{y} \parallel \hat{y}_{\text{eye}}), \tag{16}$$

where $D_{\text{KL}}(p \parallel q)$ denotes the Kullback–Leibler divergence between prediction distributions p and q.

To further enhance the discriminative capacity of specific features and strengthen feature extractors' mining of single-modality discriminative information, we incorporate strong classification loss to directly supervise f_{eeg}^{P} and f_{eye}^{P}, guiding the extractors to capture effective emotional information in single-modality features:

$$\mathcal{L}_{\text{CE}} = \text{CE}(\hat{y}_{\text{eeg}}, y) + \text{CE}(\hat{y}_{\text{eye}}, y), \tag{17}$$

where $\text{CE}(\hat{y}, y)$ denotes the cross-entropy loss between the predicted label $\hat{y}$ and the ground-truth label y.

The final MD2SC Loss is formulated as:

$$\mathcal{L}_{\text{MD2SC}} = \mathcal{L}_{\text{KL}} + \mathcal{L}_{\text{CE}}. \tag{18}$$

This joint loss design effectively enhances the representational capacity of single-modality specific features, strengthens the mining depth of feature extractors for single-modality information, and consequently significantly improves both discriminability and generalization capability of fused features.

Incorporating the aforementioned loss functions, the total training objective combines multimodal cross-entropy loss, EEGEye-CLDC Loss, and MD2SC Loss:

$$\mathcal{L} = \text{CE}(\hat{y}, y) + \lambda_1 \mathcal{L}_{\text{EEGEye-CLDC}} + \lambda_2 \mathcal{L}_{\text{MD2SC}}, \tag{19}$$

where λ_1 and λ_2 are weighting coefficients to balance the contributions of each loss term during optimization.

4 Experiments and Results

4.1 Dataset Introduction

To validate the effectiveness of the proposed model, we conduct experimental evaluations on two public multimodal EEG emotion datasets: SEED and SEED-V.

SEED: This study employs the multimodal version of the SEED dataset containing synchronized EEG and eye movement signals. The dataset uses 15 film clips (corresponding to three emotions: happiness, neutral, and sadness) as experimental stimuli. Each subject participates in three experimental sessions. Our experiments utilize complete data from all 12 subjects across three sessions, comprising 36 sessions with synchronized EEG and eye movement recordings. For each subject in the SEED dataset, we chronologically divide the 15 trials into the first 9 trials for training and the remaining 6 for testing. To ensure statistical robustness, we compute the mean and standard deviation of results across all three sessions for each subject. This experimental protocol is consistent with baseline methods such as DCCA-AM [53].

SEED-V: This dataset contains EEG and eye movement signals for five emotions (happiness, sadness, neutral, fear, disgust). The experiment requires 16 subjects (6 males, 10 females) to watch 15 film clips (3 clips per emotion), with each subject completing three experimental sessions. For each subject, we chronologically divide their 15 trials into the first 10 trials as the training set and the remaining 5 trials as the test set. The final results are reported as the mean and standard deviation across all 16 subjects to assess model robustness. This experimental protocol is consistent with that used in baseline methods such as BDAE and G2G.

4.2 Implementation Details

All experiments were conducted on a workstation equipped with NVIDIA GeForce RTX 4090 GPUs using PyTorch 2.2.2 and CUDA 12.1. The AdamW optimizer was adopted for model optimization, with the initial learning rates set to 0.008 and 0.0085 for the SEED and SEED-V datasets, respectively. The batch size was set to 32, and the maximum number of training epochs was 500. Early stopping was employed to alleviate overfitting. Except for the learning rate, all remaining hyperparameter settings were kept identical across the two datasets, and were determined through preliminary experiments. During training, a learning rate scheduling strategy combining linear warmup and MultiStep decay was adopted, where the learning rate was gradually increased during the warmup stage and decayed to 0.1 times the initial value after one-third of the total training epochs. In the EEGEye-CLDC Loss, λ_{mmd} and λ_{dec} were both fixed at 1, while the temperature coefficient τ was set to 0.07. In addition, the loss balancing coefficients λ_1 and λ_2 were both set to 1 based on sensitivity analysis on the validation set, which provided the best trade-off between classification supervision and cross-modal distribution alignment.

For the EEG branch, the feature extractor consists of a time-frequency-spatial attention module followed by a three-layer two-dimensional convolutional network. The first convolutional layer adopts 12 input channels and 64 output channels with a kernel size of 3×3, stride of 1, and padding of 1, followed by batch normalization, ReLU activation, 2×2 max pooling, and dropout with a rate of 0.3. The output channels of the second and third convolutional layers are increased to 128 and 256, respectively, while maintaining the same convolutional configuration, and no residual connections are introduced in the convolutional module. The multi-head attention module employs $H = 6$ attention heads with head embedding dimension $D_1 = 10$. The selection of H was guided by prior transformer-based EEG studies together with preliminary comparative evaluation using $H = \{4, 6, 8\}$. Considering both representation capability and computational efficiency, $H = 6$ was adopted in the final model configuration.

For the eye movement branch, a four-layer multilayer perceptron was employed to extract modality-specific representations from the 33-dimensional handcrafted eye movement features. To determine an appropriate network depth, comparative experiments were conducted using MLP architectures with 3, 4, 5, and 6 layers. The experimental results indicated that the four-layer configuration achieved the best trade-off between feature representation capability and model complexity. Therefore, a four-layer MLP was adopted as the final architecture for the eye movement modality. The embedding feature dimension was set to $D_2 = 512$.

4.3 Multimodal Emotion Recognition

To evaluate the performance of the proposed MS^{2}FL framework, we conducted comprehensive experiments on the SEED and SEED-V datasets, comparing with multiple representative baseline methods. Reported standard deviations (STD) correspond to variability across subjects, reflecting inter-subject differences. Statistical significance of performance improvements was assessed using paired t-tests with a significance level of $\alpha = 0.05$, and the resulting P-values indicate whether the observed gains are statistically meaningful. For baseline methods marked with *, results were strictly

Table 1. Comparison with baseline methods on the SEED dataset. Accuracy (ACC) and standard deviation (STD) are reported. * denotes results reproduced using official open-source codes.

Method	ACC (%)	STD
Concat	82.43	13.76
MAX	80.81	13.78
Fuzzy	84.22	12.08
BDAE (2019)	90.58	10.26
DCCA-AM (2022)	92.79	8.21
MAET (2023)*	89.90	9.57
G2G (2023)*	91.78	7.37
Ours	**96.99**	**3.54**

reproduced using the official open-source codes, following the original data splits and default hyperparameter settings to ensure a fair comparison.

4.3.1 SEED. As shown in Table 1, experimental results demonstrate that our proposed MS^{2}FL framework achieves significant performance improvements on the SEED dataset, attaining the highest accuracy of 96.99%. It surpasses the current best-performing method DCCA-AM by 4.2% (92.79%) and reduces the standard deviation to 3.54%, thereby exhibiting superior accuracy and stability compared to existing mainstream approaches. In contrast, traditional methods such as feature concatenation (Concat)[53] and max pooling fusion (MAX)[53] fail to effectively model modality complementarity, with accuracies of only 82.43% and 80.81%, respectively, and relatively high standard deviations—indicating limited generalization ability. The Fuzzy method improves fusion adaptability via fuzzy feature mapping, raising accuracy to 84.22%, yet still lacks sufficient alignment between modality-specific feature distributions. BDAE employs an autoencoder network to enhance inter-modal information interaction, increasing the accuracy to 90.58%; however, its inability to align multimodal features constrains fusion performance. MAET [54] leverages an adaptive Transformer architecture to learn multimodal information and achieves an accuracy of 89.90%; however, it also lacks a feature alignment mechanism, resulting in limited multimodal consistency. DCCA-AM integrates DCCA and attention mechanisms to enhance the alignment between EEG and eye movement features, achieving 92.79% accuracy, thereby outperforming BDAE and MAET. Nonetheless, its fusion strategy still suffers from redundant information and fails to fully exploit modality complementarity. G2G [55] adopts a graph neural network-based fusion strategy and makes progress in modeling inter-modal relationships, achieving 91.78% accuracy. However, it does not effectively balance unimodal feature integrity with multimodal interaction. To ensure the scientific rigor of the comparison, we performed statistical hypothesis testing with the best-performing open source baseline G2G, and the P-value of the paired t-test was 0.013, confirming the significance of our performance improvement.

4.3.2 SEED-V. To further evaluate the method's applicability in more challenging scenarios, we conduct additional experiments on the five-class SEED-V dataset. As shown in Table 2, our method achieves an accuracy of 89.83% on the SEED-V dataset, outperforming the current best-performing method G2G (86.24%) by 3.59 percentage points, while maintaining a relatively low standard deviation of 6.94%. Meanwhile, the early method BDAE achieves an accuracy of 79.70% on the SEED-V dataset. It mainly relies on autoencoders for feature learning but performs suboptimally when handling complex emotion categories. DGCCA [56] enhances cross-modal alignment through deep generalized canonical

Table 2. Comparison with baseline methods on the SEED-V dataset. Accuracy (ACC) and standard deviation (STD) are reported. * denotes results reproduced using official open-source codes.

Method	ACC (%)	STD
BDAE (2019)	79.70	4.76
DGCCA (2020)	82.11	**2.76**
DCCA (2022)	85.30	5.60
MAET (2023)*	80.98	8.36
G2G (2023)	86.24	7.28
MMV-SSTMA (2024)	81.38	8.76
DTCA (2024)	86.05	6.10
Ours	**89.83**	6.94

Table 3. Ablation results of different MS²FL configurations on the SEED and SEED-V datasets. Accuracy (ACC), standard deviation (STD), F1-score (F1), and recall are reported. "✓" indicates that the corresponding component is enabled.

Module		SEED				SEED-V			
SP-FDA	FDEF	ACC (%)	STD	F1 (%)	Recall (%)	ACC (%)	STD	F1 (%)	Recall (%)
		92.14	6.30	91.92	92.04	83.57	10.97	82.67	82.48
	✓	94.72	4.27	94.21	94.57	85.78	8.72	84.83	85.25
✓		94.60	5.39	94.05	94.48	86.12	9.64	85.24	85.97
✓	✓	**96.99**	**3.54**	**96.92**	**96.84**	**89.83**	**6.94**	**88.18**	**89.19**

correlation analysis with attention mechanisms, reaching an accuracy of 82.11%. However, it fails to fully exploit fine-grained information, resulting in limited performance improvement. DCCA combines deep canonical correlation analysis with attention, pushing the accuracy to 85.30%, though its feature fusion strategy remains insufficient, hindering further gains. MAET does not include explicit feature alignment, achieving an accuracy of 80.98% with a relatively high standard deviation of 8.36%. G2G leverages a graph neural network-based fusion strategy, achieving an accuracy of 86.24%. MMV-SSTMA [57] adopts a self-supervised multimodal multi-view spectro-spatial-temporal masked autoencoder and utilizes modality complementarity to achieve 81.38% accuracy, but its high standard deviation (8.76%) indicates insufficient exploitation of complementary information. DTCA [58] integrates a dual-branch Transformer with cross-attention mechanisms to aggregate enhanced information, reaching 86.05% accuracy with a reduced standard deviation of 6.10%, thus improving model stability. Similarly, we performed statistical hypothesis testing with the best-performing open source baseline G2G, and the P-value of the paired t-test was 0.0143, confirming the significance of our performance improvement.

4.4 Ablation Studies

4.4.1 Module Ablation Experiment. To further validate the effectiveness of individual components in MS²FL, we conduct systematic ablation studies on both SEED and SEED-V datasets.We first conduct a Module Ablation Experiment, and the results in Table 3 demonstrate that both modules significantly improve the accuracy and stability of multimodal emotion recognition, and their combination yields further improvements in overall performance.

Table 4. Comparison of different modality configurations on the SEED and SEED-V datasets. Accuracy (ACC), standard deviation (STD), F1-score (F1), and recall are reported for EEG-only, eye-only, and EEG+EYE fusion models. EEG indicates using only EEG signals, EYE indicates using only eye movement signals, and EEG+EYE indicates multimodal fusion of EEG and eye movement features.

Modal	SEED				SEED-V			
	ACC (%)	STD	F1 (%)	Recall (%)	ACC (%)	STD	F1 (%)	Recall (%)
EEG	89.69	8.68	88.70	89.39	81.97	11.13	80.84	81.19
EYE	87.13	9.71	86.53	86.92	72.82	7.97	68.91	70.91
EEG+EYE	**96.99**	**3.54**	**96.92**	**96.84**	**89.83**	**6.94**	**88.18**	**89.19**

On the SEED dataset, the baseline model (without SP-FDA and FDEF) achieved an accuracy of 92.14% with a standard deviation of 6.30%. The corresponding F1-score and recall also indicate that the model still suffers from certain classification instability. When only FDEF is introduced, the accuracy increases to 94.72% and the standard deviation decreases to 4.27%, while the F1-score and recall are also improved, demonstrating that enhancing the peak-valley differences of discriminative features and introducing residual mechanisms can effectively capture cross-modal complementary information and improve the discriminative power of fused features. Similarly, introducing only SP-FDA also improves the accuracy to 94.60% with a standard deviation of 5.39%, accompanied by improvements in F1-score and recall, verifying the effectiveness of aligning cross-modal feature distributions while strengthening unimodal independent representation. Finally, when both SP-FDA and FDEF are combined, the model achieves a further accuracy improvement to 96.99%, and the standard deviation significantly drops to 3.54%. Meanwhile, the F1-score and recall are further improved, indicating substantial improvements in both classification performance and stability.

On the SEED-V dataset, the overall trend remains consistent with that observed on SEED. However, due to the increased difficulty of the task, the baseline model only achieves an accuracy of 83.57% with a standard deviation of 10.97%, while the F1-score and recall indicate greater performance fluctuations. After introducing FDEF, the accuracy increases to 85.78% and the standard deviation decreases to 8.72%, with corresponding improvements in F1-score and recall. Similarly, incorporating SP-FDA results in an accuracy of 86.12% with a reduced standard deviation of 9.64%, further improving the F1-score and recall metrics. These results respectively highlight the effectiveness of feature enhancement and distribution consistency optimization in handling complex emotion recognition tasks. When both SP-FDA and FDEF are introduced, the accuracy further improves to 89.83% and the standard deviation significantly decreases to 6.94%, while the F1-score and recall are also improved, demonstrating substantial improvements in both discriminative ability and model stability.

4.4.2 Modal Ablation Experiment. To comprehensively evaluate the effectiveness of the proposed multimodal fusion framework MS^{2}FL in emotion recognition tasks, modal ablation experiments were conducted on SEED and SEED-V. By integrating the overall performance comparison results presented in Table 4 with the detailed analysis of the confusion matrices in Fig. 3, the superiority of the fusion method in improving classification accuracy, stability, and reducing inter-class confusion is systematically demonstrated.

On the SEED dataset, Table 4 shows that using EEG data alone achieves an emotion classification accuracy of 89.69%, with the F1-score and recall indicating slightly lower performance. Using eye movement data alone achieves 87.13% accuracy, with F1-score and recall following a similar trend. After fusing EEG and eye movement data, accuracy significantly improves to 96.99%, while the F1-score and recall are also increased, and the standard deviation decreases from 8.68% and 9.71% in unimodal cases to 3.54%. Analysis of confusion matrices in Fig. 3(a-c) reveals clear

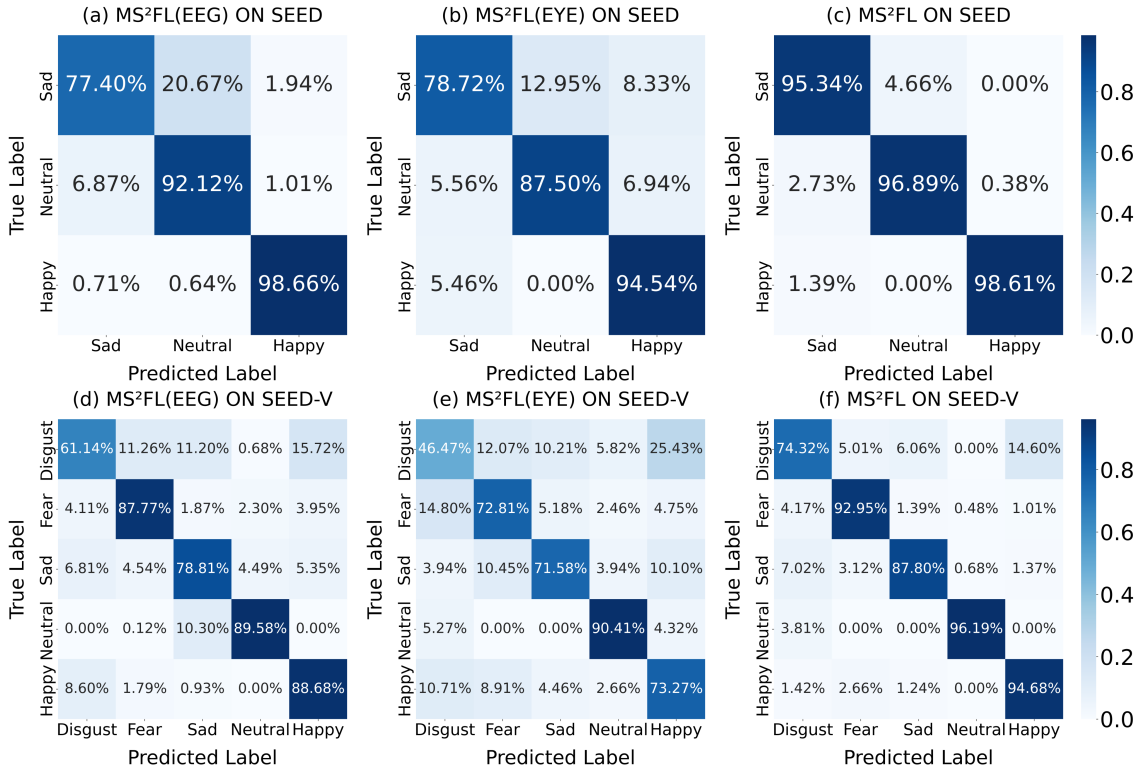

Fig. 3. Confusion matrices of MS²FL using different data sources on the SEED and SEED-V datasets. Panels (a), (b), and (c) show classification results on the SEED dataset using EEG data, eye movement data, and fused data, respectively. Panels (d), (e), and (f) show classification results on the SEED-V dataset using EEG data, eye movement data, and fused data, respectively.

complementarity between EEG and eye movement data in classifying Neutral and Sad emotions. In the unimodal EEG model, Neutral is correctly classified in 92.12% of cases, while Sad is correctly classified in 77.40% and misclassified mainly as Neutral with a rate of 20.67%. In the unimodal eye movement model, Sad recognition improves to 78.72%, but Neutral accuracy slightly decreases to 87.50%. After fusion in MS²FL, Neutral accuracy increases to 96.89% and Sad recognition to 95.34%, demonstrating a substantial reduction in confusion between these two categories. Happy emotion is accurately recognized in all models, maintaining high accuracy after fusion at 98.61%, comparable to the unimodal models with 98.66% for EEG and 94.54% for eye movement. These results provide quantitative evidence that the fusion strategy enhances class-level discriminability, particularly for emotion pairs that are prone to mutual confusion.

On the SEED-V dataset, the results similarly confirm the benefits of the fusion strategy. As shown in Table 4, using EEG data alone achieves an accuracy of 81.97%, with F1-score and recall showing lower performance, while eye movement alone reaches 72.82% accuracy, with corresponding F1-score and recall also lower. In contrast, MS²FL significantly improves accuracy to 89.83%, while the F1-score and recall are also enhanced, and the standard deviation decreases from 11.13% and 7.97% in unimodal cases to 6.94%, indicating enhanced performance stability. Detailed analysis of the confusion matrices in Fig. 3(d-f) shows that fusion improves recognition for specific categories. Disgust is recognized with 61.14% accuracy using EEG and 46.47% using eye movement, and MS²FL raises this to 74.32%,

Table 5. Ablation results of different loss component configurations on the SEED and SEED-V datasets. Accuracy (ACC), standard deviation (STD), F1-score (F1), and recall are reported. "✓" indicates that the corresponding loss component is included in the optimization objective.

Loss			SEED				SEED-V			
$\mathcal{L}_{\mathbf{mmd}}$	$\mathcal{L}_{\mathbf{dec}}$	$\mathcal{L}_{\mathbf{KL}}$	**ACC (%)**	**STD**	**F1 (%)**	**Recall (%)**	**ACC (%)**	**STD**	**F1 (%)**	**Recall (%)**
	✓	✓	93.49	5.44	93.37	93.53	85.26	9.50	83.81	87.40
✓		✓	94.27	4.42	94.21	94.44	86.22	7.62	84.40	85.33
✓	✓		94.36	5.46	94.19	94.09	86.38	8.22	85.28	84.47
✓	✓	✓	**96.99**	**3.54**	**96.92**	**96.84**	**89.83**	**6.94**	**88.18**	**89.19**

notably reducing confusion with Fear and Sad. Fear achieves 87.77% accuracy with EEG and 72.81% with eye movement, while MS^{2}FL improves it to 92.95%, mitigating misclassification as Disgust and Sad. For Sad, unimodal accuracies are 78.81% and 71.58%, and MS^{2}FL increases it to 87.80%, reducing confusion with Disgust and Fear. Neutral classification improves from 89.58% (EEG) and 90.41% (EYE) to 96.19% after fusion, while Happy recognition reaches 94.68%, correcting most misclassifications present in the eye movement model. These quantitative results confirm that fusion enhances class-level discriminability and reduces confusion between closely related emotions.

4.4.3 Loss Ablation Experiment. To further investigate the contributions of different loss components, additional ablation experiments were conducted by individually removing $\mathcal{L}_{\text{mmd}}$, $\mathcal{L}_{\text{dec}}$, and $\mathcal{L}_{\text{KL}}$ from the optimization objective. The experimental results on the SEED and SEED-V datasets are presented in Table 5.

As shown in Table 5, the complete model achieves the best overall performance on both datasets, indicating that each loss component contributes positively to the optimization process. On the SEED dataset, removing $\mathcal{L}_{\text{mmd}}$ decreases the accuracy from 96.99% to 93.49%, while the F1-score and recall also decline accordingly. Meanwhile, the standard deviation increases from 3.54% to 5.44%, suggesting that the distribution alignment constraint contributes to the consistency and stability of shared feature representations across modalities. When $\mathcal{L}_{\text{dec}}$ is removed, the accuracy decreases to 94.27%, while the F1-score and recall also show a corresponding decline, indicating that the decoupling constraint contributes to preserving modality-specific discriminative information. Similarly, removing $\mathcal{L}_{\text{KL}}$ results in an accuracy of 94.36%, accompanied by reductions in F1-score and recall, reflecting the contribution of cross-modal knowledge transfer during feature optimization.

A similar trend can also be observed on the SEED-V dataset. Specifically, removing $\mathcal{L}_{\text{mmd}}$ reduces the accuracy from 89.83% to 85.26%, while the F1-score and recall also decrease accordingly. Without $\mathcal{L}_{\text{dec}}$, the model achieves an accuracy of 86.22%, and removing $\mathcal{L}_{\text{KL}}$ yields an accuracy of 86.38%. Across all three ablation settings, the F1-score, recall, and standard deviation metrics also decline compared with the complete model. Overall, the results indicate that the different loss components jointly contribute to feature alignment, modality-specific representation preservation, and cross-modal feature interaction, thereby improving classification performance and model stability.

4.5 Feature Visualization

As shown in Fig. 4, to intuitively evaluate the effectiveness of MS^{2}FL in enhancing feature discriminability, the t-SNE [59] algorithm was employed to visualize the distributions of six features from a representative subject in the SEED and SEED-V datasets. In SEED (Fig. 4(a)), EEG features form loose clusters with noticeable overlap, especially between sad and neutral, while eye-movement features appear even more scattered, offering little category separation. Baseline

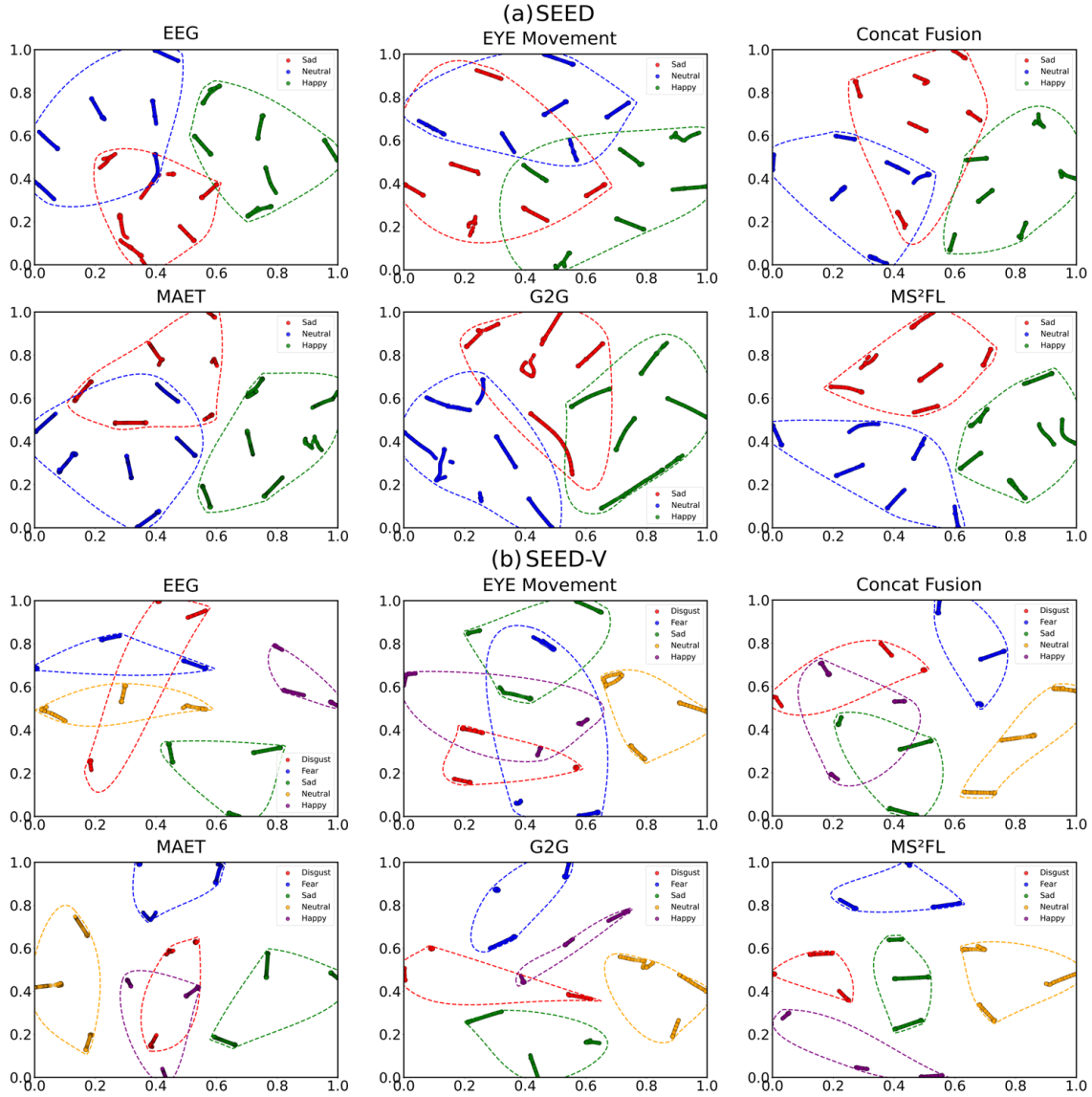

Fig. 4. Visualization of feature distributions under different methods for representative subjects in the two datasets (a) SEED and (b) SEED-V. For each dataset, the visualization corresponds in turn to EEG, eye movements, fused features from a representative baseline fusion method, and fused features obtained from MS²FL.

fusion methods tighten the distribution somewhat, yet many samples remain mingled at class boundaries. By contrast, MS²FL produces compact, well-defined clusters with clear margins, indicating markedly improved intra-class cohesion and inter-class separation. The same pattern emerges on SEED-V (Fig. 4(b)): EEG and eye-movement features alone show widespread overlap among emotions such as disgust, sad, and fear; baseline fusion yields only modest gains; but

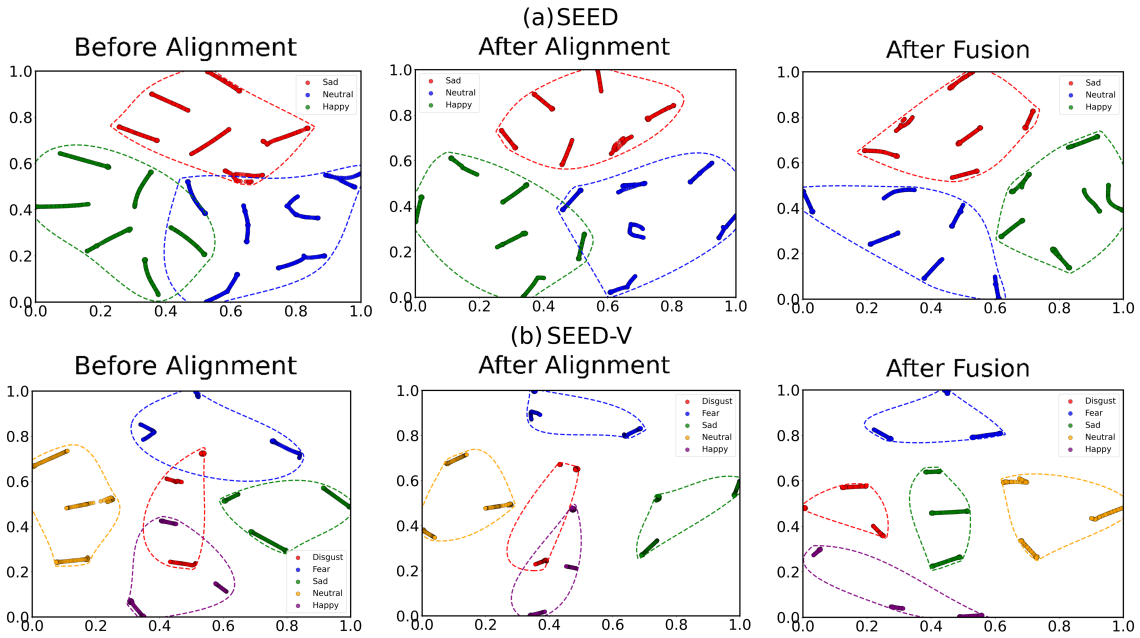

Fig. 5. Visualization of feature distributions at different stages for representative subjects in the two datasets (a) SEED and (b) SEED-V. For each dataset, the visualizations correspond to the feature distributions before alignment, after alignment, and after fusion, respectively.

MS^{2}FL achieves visibly aligned, discriminative clusters, substantially boosting the separability of complex emotional states.

To further illustrate how SP-FDA reshapes the shared feature space, additional t-SNE visualizations are presented in Fig. 5. Unlike the comparative visualization in Fig. 4, Fig. 5 shows the feature distributions before alignment, after SP-FDA alignment, and after multimodal fusion. As shown in Fig. 5(a), the SEED dataset exhibits noticeable dispersion and overlap before alignment, while SP-FDA makes intra-class samples more compact and inter-class boundaries clearer. After fusion, the clusters become further separated. Similar trends can be observed on the SEED-V dataset in Fig. 5(b), where SP-FDA effectively reduces inter-class mixing and improves the consistency of the shared feature space, thereby enhancing feature discriminability for subsequent multimodal fusion and emotion classification.

5 Discussion

Multimodal emotion recognition is a key technology in human-computer interaction. However, existing approaches still face challenges in simultaneously ensuring cross-modal consistency and preserving modality-specific expressiveness. To address this issue, we propose a novel MS^{2}FL framework, which aims to model both modality complementarity and modality specificity. The framework incorporates two core modules: SP-FDA and FDEF. SP-FDA aligns shared features across modalities while mitigating distributional heterogeneity and retaining the independent characteristics of EEG and eye movement signals. FDEF enhances the discriminability of fused features by emphasizing informative variations and effectively capturing complementary information. By integrating these two modules, MS^{2}FL simultaneously preserves

unimodal feature independence and exploits modality complementarity. This unified design explains the consistent performance gains observed in our experiments and lays a solid foundation for the subsequent discussion.

5.1 Comparative Analysis

In the comparison methods, the simple fusion approaches Concat, MAX, and Fuzzy fail to effectively capture cross-modal complementarities, resulting in lower accuracy and higher variance. BDAE, which leverages autoencoder-based cross-modal interactions, and MAET, employing an adaptive Transformer for multimodal representation learning, achieve moderate accuracy improvements of approximately 5–6% over traditional methods, yet both lack feature alignment mechanisms, limiting multi-modal feature distribution consistency and fusion performance. DCCA and its extension DCCA-AM further enhance the alignment between EEG and eye movement features, outperforming BDAE and MAET with accuracy gains of roughly 2–5% and reaching 92.79% on SEED. However, their fusion strategies are overly simple and fail to fully exploit modality-specific cues and complementary information. The G2G framework adopts a graph neural network–based fusion strategy, achieving 86.24% on SEED-V, but it does not effectively balance unimodal feature characteristics with shared multimodal representations. In contrast, MS^{2}FL explicitly addresses these limitations through SP-FDA and FDEF, preserving modality-specific information while enhancing inter-modal consistency and feature discriminability. As a result, MS^{2}FL achieves superior recognition performance, yielding an additional 3–4% improvement over the mainstream methods across both SEED and SEED-V and attaining the highest overall accuracy.

5.2 Ablation Study Analysis

The ablation studies collectively demonstrate the effectiveness of the proposed MS^{2}FL framework from the perspectives of module design, modality fusion, and optimization objectives. In the module-level experiments, both SP-FDA and FDEF consistently improve the performance of the baseline model on the SEED and SEED-V datasets. Specifically, SP-FDA reduces cross-modal distribution discrepancies while preserving modality-specific characteristics, thereby improving the consistency of shared feature representations. Meanwhile, FDEF enhances the discriminability of fused features by strengthening peak-valley distribution differences within the semantic space and promoting complementary information interaction between modalities. When the two modules are jointly introduced, the model achieves the best overall performance, indicating that SP-FDA and FDEF provide complementary advantages for multimodal representation learning. The modality comparison experiments further verify the necessity of multimodal fusion. Although EEG-based recognition achieves relatively stronger performance than eye movement signals, unimodal methods still exhibit noticeable confusion among similar emotional categories. After multimodal fusion, MS^{2}FL significantly improves classification accuracy and reduces category-level confusion, as reflected by the more concentrated main diagonal and fewer misclassifications in the confusion matrices. Furthermore, the loss ablation experiments show that removing $\mathcal{L}_{\text{mmd}}$, $\mathcal{L}_{\text{dec}}$, or $\mathcal{L}_{\text{KL}}$ all leads to performance degradation in terms of accuracy, F1-score, recall, and stability. In particular, the larger decline caused by removing $\mathcal{L}_{\text{mmd}}$ indicates the importance of reducing distribution discrepancies among shared features, while the degradation observed after removing $\mathcal{L}_{\text{dec}}$ and $\mathcal{L}_{\text{KL}}$ suggests that modality-specific representation preservation and cross-modal knowledge interaction both contribute to effective feature optimization. Overall, these results confirm that the different modules and optimization objectives within MS^{2}FL collaboratively improve the consistency, discriminability, and stability of multimodal emotional representations.

5.3 Feature Visualization Analysis

Feature visualization provides compelling evidence for the discriminative capacity of the proposed MS^{2}FL framework. In unimodal representations, EEG features form loosely clustered distributions with substantial overlap, particularly between closely related emotions such as Sad and Neutral, while eye-movement features exhibit even greater dispersion and limited class separability. Conventional fusion methods achieve modest improvements in cluster cohesion, yet many samples remain ambiguously distributed near class boundaries. By contrast, MS^{2}FL generates compact, well-separated clusters with clear margins, indicating enhanced intra-class consistency and inter-class discriminability. Additional visualizations further show that SP-FDA effectively reduces feature overlap and progressively improves cluster compactness during the alignment and fusion process. These findings suggest that the framework effectively reduces misclassification among ambiguous categories, an advantage attributable to its ability to exploit the complementary characteristics of EEG and eye-movement signals while preserving modality-specific information in the fused representations. The improved separability of emotional states observed in both SEED and SEED-V datasets underscores the capacity of MS^{2}FL to support precise emotion recognition in complex, multimodal affective computing scenarios.

5.4 Limitations and Future Work

Despite the promising performance of MS^{2}FL, several limitations remain. In some experiments, individual subjects exhibited suboptimal results, and the framework's performance on external datasets was lower than expected, highlighting challenges in cross-subject and cross-dataset generalization. These limitations may arise from low-quality or noisy EEG and eye-movement signals, inherent inter-subject variability, and dataset-specific characteristics. Furthermore, while the current framework effectively captures complementary information and preserves modality-specific features, its adaptability to highly diverse real-world conditions has not been fully tested. Future work will focus on enhancing the robustness of MS^{2}FL across subjects and datasets by exploring adaptive fusion strategies, dynamic weighting of modalities, and domain adaptation techniques. In addition, integrating real-time feedback mechanisms and lightweight model implementations for edge deployment could further extend the practical applicability of the framework in clinical assessment, cognitive workload evaluation, and affective computing tasks in real-world environments.

6 Conclusion

This paper presents a novel multimodal emotion recognition framework named MS^{2}FL, which adopts an alignment-then-fusion strategy to address the challenge of simultaneously modeling modality complementarity and modality specificity. It employs SP-FDA to enforce feature distribution consistency for modality-shared features while preserving modality-specific features, and utilizes FDEF to achieve effective fusion by enhancing the discriminability of these features.Through extensive experiments, MS^{2}FL consistently outperforms both state-of-the-art methods and traditional baselines in accuracy and stability.Ablation studies and modal comparison experiments validate the effectiveness of its alignment and fusion mechanisms in leveraging multimodal complementarities.Overall, MS^{2}FL offers a robust and scalable approach for practical multimodal emotion recognition, with significant implications for enhancing affective computing and human–machine interaction applications.

Acknowledgments

This work is supported by the Natural Science Research Project of Anhui Educational Committee under Grant (No. 2024AH050054), Distinguished Youth Foundation of Anhui Scientific Committee (No. 2208085J05), National Natural Science Foundation of China (NSFC) (Nos. 62476004, 62501007), Cloud Ginger XR-1 platform.

References

[1] Kaouther Ezzameli and Hela Mahersia. Emotion recognition from unimodal to multimodal analysis: A review. *Information Fusion*, 99:101847, 2023.

[2] Matteo Spezialetti, Giuseppe Placidi, and Silvia Rossi. Emotion recognition for human-robot interaction: Recent advances and future perspectives. *Frontiers in Robotics and AI*, 7:532279, 2020.

[3] Weizhi Meng, Yong Cai, Laurence T Yang, and Wei-Yang Chiu. Hybrid emotion-aware monitoring system based on brainwaves for internet of medical things. *IEEE Internet of Things Journal*, 8(21):16014–16022, 2021.

[4] Klaus-Robert Müller, Michael Tangermann, Guido Dornhege, Matthias Krauledat, Gabriel Curio, and Benjamin Blankertz. Machine learning for real-time single-trial eeg-analysis: from brain–computer interfacing to mental state monitoring. *Journal of neuroscience methods*, 167(1):82–90, 2008.

[5] Guanglong Du, Jinshao Su, Linlin Zhang, Kang Su, Xueqian Wang, Shaohua Teng, and Peter Xiaoping Liu. A multi-dimensional graph convolution network for eeg emotion recognition. *IEEE Transactions on Instrumentation and Measurement*, 71:1–11, 2022.

[6] Beatriz García-Martínez, Arturo Martinez-Rodrigo, Raul Alcaraz, and Antonio Fernández-Caballero. A review on nonlinear methods using electroencephalographic recordings for emotion recognition. *IEEE Transactions on Affective Computing*, 12(3):801–820, 2019.

[7] Jindi Bao, Xiaomei Tao, and Yinghui Zhou. An emotion recognition method based on eye movement and audiovisual features in mooc learning environment. *IEEE Transactions on Computational Social Systems*, 11(1):171–183, 2022.

[8] Hodam Kim, Dan Zhang, Laehyun Kim, and Chang-Hwan Im. Classification of individual's discrete emotions reflected in facial microexpressions using electroencephalogram and facial electromyogram. *Expert Systems with Applications*, 188:116101, 2022.

[9] Pamela Zontone, Antonio Affanni, Riccardo Bernardini, Alessandro Piras, Roberto Rinaldo, Fabio Formaggia, Diego Minen, Michela Minen, and Carlo Savorgnan. Car driver's sympathetic reaction detection through electrodermal activity and electrocardiogram measurements. *IEEE Transactions on Biomedical Engineering*, 67(12):3413–3424, 2020.

[10] Ke Zhang, Yuanqing Li, Jingyu Wang, Erik Cambria, and Xuelong Li. Real-time video emotion recognition based on reinforcement learning and domain knowledge. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3):1034–1047, 2021.

[11] Siddique Latif, Rajib Rana, Sara Khalifa, Raja Jurdak, Junaid Qadir, and Björn Schuller. Survey of deep representation learning for speech emotion recognition. *IEEE Transactions on Affective Computing*, 14(2):1634–1654, 2021.

[12] Jiawen Deng and Fuji Ren. A survey of textual emotion recognition and its challenges. *IEEE Transactions on Affective Computing*, 14(1):49–67, 2021.

[13] Fernando Lopes da Silva. Eeg and meg: relevance to neuroscience. *Neuron*, 80(5):1112–1128, 2013.

[14] Soraia M Alarcao and Manuel J Fonseca. Emotions recognition using eeg signals: A survey. *IEEE transactions on affective computing*, 10(3):374–393, 2017.

[15] Matti Hämäläinen, Riitta Hari, Risto J Ilmoniemi, Jukka Knuutila, and Olli V Lounasmaa. Magnetoencephalography—theory, instrumentation, and applications to noninvasive studies of the working human brain. *Reviews of modern Physics*, 65(2):413, 1993.

[16] Claudio Aracena, Sebastián Basterrech, Václav Snáel, and Juan Velásquez. Neural networks for emotion recognition based on eye tracking data. In *2015 IEEE International Conference on Systems, Man, and Cybernetics*, pages 2632–2637. IEEE, 2015.

[17] Yiting Wang, Jia-Wen Liu, Bao-Liang Lu, and Wei-Long Zheng. From eeg to eye movements: Cross-modal emotion recognition using constrained adversarial network with dual attention. *IEEE Transactions on Affective Computing*, 2024.

[18] Wei-Long Zheng and Bao-Liang Lu. Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks. *IEEE Transactions on autonomous mental development*, 7(3):162–175, 2015.

[19] Wei Liu, Jie-Lin Qiu, Wei-Long Zheng, and Bao-Liang Lu. Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 14(2):715–729, 2021.

[20] Xiaobing Du, Cuixia Ma, Guanhua Zhang, Jinyao Li, Yu-Kun Lai, Guozhen Zhao, Xiaoming Deng, Yong-Jin Liu, and Hongan Wang. An efficient lstm network for emotion recognition from multichannel eeg signals. *IEEE Transactions on Affective Computing*, 13(3):1528–1540, 2020.

[21] Wei Liu, Jie-Lin Qiu, Wei-Long Zheng, and Bao-Liang Lu. Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 14(2):715–729, 2021.

[22] Wei Liu, Wei-Long Zheng, and Bao-Liang Lu. Emotion recognition using multimodal deep learning. In *International conference on neural information processing*, pages 521–529. Springer, 2016.

[23] Yong Zhang, Cheng Cheng, Shuai Wang, and Tianqi Xia. Emotion recognition using heterogeneous convolutional neural networks combined with multimodal factorized bilinear pooling. *Biomedical Signal Processing and Control*, 77:103877, 2022.

[24] Minghai Chen, Sen Wang, Paul Pu Liang, Tadas Baltrušaitis, Amir Zadeh, and Louis-Philippe Morency. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In *Proceedings of the 19th ACM international conference on multimodal interaction*, pages 163–171, 2017.

[25] Haiyang Xu, Hui Zhang, Kun Han, Yun Wang, Yiping Peng, and Xiangang Li. Learning alignment for multimodal emotion recognition from speech. *arXiv preprint arXiv:1909.05645*, 2019.

[26] Wei-Long Zheng, Jia-Yi Zhu, and Bao-Liang Lu. Identifying stable patterns over time for emotion recognition from eeg. *IEEE transactions on affective computing*, 10(3):417–429, 2017.

[27] Qi Li, Yunqing Liu, Fei Yan, Qiong Zhang, and Cong Liu. Emotion recognition based on multiple physiological signals. *Biomedical Signal Processing and Control*, 85:104989, 2023.

[28] Hao Chao, Liang Dong, Yongli Liu, and Baoyun Lu. Emotion recognition from multiband eeg signals using capsnet. *Sensors*, 19(9):2212, 2019.

[29] TN Kipf. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[30] Nazmi Sofian Suhaimi, James Mountstephens, and Jason Teo. Eeg-based emotion recognition: a state-of-the-art review of current trends and opportunities. *Computational intelligence and neuroscience*, 2020(1):8875426, 2020.

[31] Tengfei Song, Wenming Zheng, Peng Song, and Zhen Cui. Eeg emotion recognition using dynamical graph convolutional neural networks. *IEEE Transactions on Affective Computing*, 11(3):532–541, 2018.

[32] Xinhui Li, Guowang Zhuang, Minchao Wu, and Zhao Lv. Eeg correlation analysis-guided graph local enhanced feature learning for emotion recognition. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[33] AV Geetha, T Mala, D Priyanka, and E Uma. Multimodal emotion recognition with deep learning: advancements, challenges, and future directions. *Information Fusion*, 105:102218, 2024.

[34] Hailun Lian, Cheng Lu, Sunan Li, Yan Zhao, Chuangao Tang, and Yuan Zong. A survey of deep learning-based multimodal emotion recognition: Speech, text, and face. *Entropy*, 25(10):1440, 2023.

[35] MaoSong Yan, Zhen Deng, BingWei He, ChengSheng Zou, Jie Wu, and ZhaoJu Zhu. Emotion classification with multichannel physiological signals using hybrid feature and adaptive decision fusion. *Biomedical Signal Processing and Control*, 71:103235, 2022.

[36] Xinda Li. Tacoformer: Token-channel compounded cross attention for multimodal emotion recognition. *arXiv preprint arXiv:2306.13592*, 2023.

[37] Yinsai Guo, Hang Yu, Liyan Ma, Liang Zeng, and Xiangfeng Luo. Thfe: A triple-hierarchy feature enhancement method for tiny boat detection. *Engineering Applications of Artificial Intelligence*, 123:106271, 2023.

[38] Jiehao Tang, Zhuang Ma, Kaiyu Gan, Jianhua Zhang, and Zhong Yin. Hierarchical multimodal-fusion of physiological signals for emotion recognition with scenario adaption and contrastive alignment. *Information Fusion*, 103:102129, 2024.

[39] Jing Li, Ning Chen, Hongqing Zhu, Guangqiang Li, Zhangyong Xu, and Dingxin Chen. Incongruity-aware multimodal physiology signals fusion for emotion recognition. *Information Fusion*, 105:102220, 2024.

[40] Yue Gu, Kangning Yang, Shiyu Fu, Shuhong Chen, Xinyu Li, and Ivan Marsic. Multimodal affective analysis using hierarchical attention strategy with word-level alignment. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2018, page 2225, 2018.

[41] Morteza Rohanian, Julian Hough, Matthew Purver, et al. Detecting depression with word-level multimodal fusion. In *Interspeech*, pages 1443–1447, 2019.

[42] Mahesh G Huddar, Sanjeev S Sannakki, and Vijay S Rajpurohit. Attention-based word-level contextual feature extraction and cross-modality fusion for sentiment analysis and emotion classification. *International Journal of Intelligent Engineering Informatics*, 8(1):1–18, 2020.

[43] Bingzhi Chen, Qi Cao, Mixiao Hou, Zheng Zhang, Guangming Lu, and David Zhang. Multimodal emotion recognition with temporal and semantic consistency. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3592–3603, 2021.

[44] Yuntao Shou, Tao Meng, Wei Ai, Fuchen Zhang, Nan Yin, and Keqin Li. Adversarial alignment and graph fusion via information bottleneck for multimodal emotion recognition in conversations. *Information Fusion*, 112:102590, 2024.

[45] Tao Meng, Fuchen Zhang, Yuntao Shou, Hongen Shao, Wei Ai, and Keqin Li. Masked graph learning with recurrent alignment for multimodal emotion recognition in conversation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[46] Yixin Wang, Shuang Qiu, Dan Li, Changde Du, Bao-Liang Lu, and Huiguang He. Multi-modal domain adaptation variational autoencoder for eeg-based emotion recognition. *IEEE/CAA Journal of Automatica Sinica*, 9(9):1612–1626, 2022.

[47] Yixin Wang, Shuang Qiu, Dan Li, Changde Du, Bao-Liang Lu, and Huiguang He. Multi-modal domain adaptation variational autoencoder for eeg-based emotion recognition. *IEEE/CAA Journal of Automatica Sinica*, 9(9):1612–1626, 2022.

[48] Ruo-Nan Duan, Jia-Yi Zhu, and Bao-Liang Lu. Differential entropy feature for eeg-based emotion classification. In *2013 6th international IEEE/EMBS conference on neural engineering (NER)*, pages 81–84. IEEE, 2013.

[49] Ed Bullmore and Olaf Sporns. The economy of brain network organization. *Nature reviews neuroscience*, 13(5):336–349, 2012.

[50] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

[51] Yifei Lu, Wei-Long Zheng, Binbin Li, and Bao-Liang Lu. Combining eye movements and eeg to enhance emotion recognition. In *IJCAI*, volume 15, pages 1170–1176. Buenos Aires, 2015.

[52] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The journal of machine learning research*, 13(1):723–773, 2012.

[53] Wei Liu, Wei-Long Zheng, Ziyi Li, Si-Yuan Wu, Lu Gan, and Bao-Liang Lu. Identifying similarities and differences in emotion recognition with eeg and eye movements among chinese, german, and french people. *Journal of Neural Engineering*, 19(2):026012, 2022.

[54] Wei-Bang Jiang, Xuan-Hao Liu, Wei-Long Zheng, and Bao-Liang Lu. Multimodal adaptive emotion transformer with flexible modality inputs on a novel dataset with continuous labels. In *proceedings of the 31st ACM international conference on multimedia*, pages 5975–5984, 2023.

[55] Ming Jin and Jinpeng Li. Graph to grid: Learning deep representations for multimodal emotion recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5985–5993, 2023.

[56] Yu-Ting Lan, Wei Liu, and Bao-Liang Lu. Multimodal emotion recognition using deep generalized canonical correlation analysis with an attention mechanism. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6. IEEE, 2020.

[57] Pengxuan Gao, Tianyu Liu, Jia-Wen Liu, Bao-Liang Lu, and Wei-Long Zheng. Multimodal multi-view spectral-spatial-temporal masked autoencoder for self-supervised emotion recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1926–1930. IEEE, 2024.

[58] Xiaoshan Zhang, Enze Shi, Sigang Yu, and Shu Zhang. Dtca: Dual-branch transformer with cross-attention for eeg and eye movement data fusion. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 141–151. Springer, 2024.

[59] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

Received 29 January 2026; revised 8 June 2026; accepted 11 July 2026