DeepEmoNet: A Lightweight Context-Aware CNN for Multimodal Emotion Recognition

Sumitra A. Jakhete
Department of Computer Science and Engineering, MIT School of Computing, MIT Art Design and Technology University, Pune, India | Department of Information Technology, SCTR's Pune Institute of Computer Technology, Pune, India
sumeetra.kasat@gmail.com (corresponding author)

Nilima Kulkarni
Department of Computer Science and Engineering, MIT School of Computing, MIT Art Design and Technology University, Pune, India
nilima.amrita@gmail.com

Received: 1 October 2025 | Revised: 16 November 2025, 13 December 2025, and 17 December 2025 | Accepted: 18 December 2025

 DOI: https://doi.org/10.48084/etasr.15258

ABSTRACT

Multimodal emotion recognition in real-world environments remains challenging due to occlusions, class imbalance, and the high computational cost of existing deep models. This paper presents DeepEmoNet, a lightweight multimodal Convolutional Neural Network (CNN) designed to integrate facial, gait, scene, and socio-dynamic depth cues through an early-fusion architecture based on Depthwise Separable Convolutions (DSCs). The model aims to achieve robust emotion recognition while maintaining low computational overhead suitable for real-time applications. Experiments on the GroupWalk dataset comprising 3,544 annotated agents across 45 environments demonstrate that DeepEmoNet achieves 91.3% accuracy and 86.5% mean Average Precision (mAP), outperforming Inception V3, ResNet-50, MobileNetV2, and recent multimodal baselines. Extended ablation studies highlight the importance of contextual modalities and early fusion, with four DSC modules offering the best accuracy–efficiency balance. Inference analysis further shows a latency of 14.8 ms/frame (~67 frames per second (FPS)), supporting real-time deployment. Overall, DeepEmoNet offers an efficient, context-aware multimodal CNN framework for emotion recognition in surveillance, smart environments, and human–computer interaction.

Keywords-affective computing; Convolutional Neural Network (CNN); computational efficiency; deep learning; Depthwise Separable Convolution (DSC); emotion recognition; multimodal fusion; real-time inference

I. INTRODUCTION

Emotions play a central role in shaping human communication, decisions, and social interaction. Automatic emotion recognition has therefore become a cornerstone of Human–Computer Interaction (HCI) with applications in healthcare, e-learning, mental health support, surveillance, and immersive technologies such as Augmented Reality/Virtual Reality (AR/VR). Early unimodal approaches based on facial expressions, speech, or text alone proved insufficient, as they overlooked the complexity and ambiguity of emotional expressions influenced by environment and social dynamics. This gave rise to Multimodal Emotion Recognition (MER), which combines facial, vocal, and contextual cues for richer understanding [1, 2]. More recently, Context-Aware Multimodal Emotion Recognition (CMER) has emerged as an advanced paradigm, embedding environmental, situational, and inter-agent cues to achieve higher robustness and ecological validity [3]. Deep learning has been instrumental in this shift: Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformer models have enabled extraction of high-level representations across modalities [3, 4], whereas attention mechanisms have improved context-sensitive fusion [5]. Nevertheless, robust CMER remains challenging due to the heterogeneity of contextual cues, the limited availability of annotated multimodal datasets, and the heavy computational demands of current deep models [5].

Several recent studies highlight both the promise and limitations of current approaches. Authors in [5] demonstrated the potential of multimodal transformers for unaligned language–vision sequences, whereas authors in [6] introduced a context-aware joint representation learning framework. Authors in [7] achieved state-of-the-art multimodal emotion analysis in conversational settings by disentangling modality and context during feature fusion, and authors in [8] explored MER in unconstrained "in-the-wild" environments. Several

studies have proposed CNN-based frameworks for efficient emotion recognition, highlighting the effectiveness of CNNs for multimodal feature extraction and fusion [9-12]. The cited self-references are limited to prior peer-reviewed works that introduced foundational preprocessing strategies, baseline fusion paradigms, or preliminary multimodal frameworks that directly informed the architectural and methodological choices of the present study.

Foundational works by authors in [13-18] emphasized the importance of contextual and multimodal integration, whereas authors in [1] provided a comprehensive survey of MER fundamentals. Several surveys have recently synthesized the rapid growth in multimodal emotion recognition, outlining trends, fusion methods, datasets, and challenges across speech, text, face, and physiological signals [18-23]. Emerging architectures such as transformer-based fusion models like HyFusER [17], hybrid EEG-Video systems [22], deep fusion of ECG and EEG [22], and gated transformer frameworks [24] are pushing accuracy and robustness within noisy or physiological contexts [24-26]. However, challenges persist as many models are designed for constrained lab datasets (e.g., IEMOCAP, CMU-MOSEI), lack rich contextual annotations, or require prohibitive computational resources. Moreover, balancing accuracy, efficiency, and generalizability in dynamic, real-world environments remain an open research problem.

Despite progress in multimodal and context-aware emotion recognition, existing models still face two central limitations. First, many architectures rely on heavy backbones (e.g., ResNet, Transformers) that are expensive to train and deploy, making them unsuitable for real-time or resource-constrained environments. Second, even state-of-the-art multimodal fusion strategies often emphasize facial or body cues while underutilizing crucial scene and socio-dynamic context, leading to misclassification when emotional expressions are subtle or ambiguous in crowded or unconstrained settings. Although significant progress has been made in deep learning-based emotion recognition, existing multimodal approaches struggle in unconstrained environments due to occlusions, limited contextual modeling, high computational complexity, and suboptimal fusion strategies. Current CNN architectures often treat modalities independently or rely on heavy networks, leading to reduced inference speed and poor generalization across socio-dynamic interactions.

Therefore, a lightweight, context-aware, and robust multimodal architecture is needed to achieve reliable group-level emotion recognition in real-world scenarios. This study offers the following novel contributions:

- A lightweight multimodal CNN (DeepEmoNet) integrating facial, gait, scene, and socio-dynamic depth cues through an efficient early-fusion pipeline.
- A Depthwise Separable Convolution (DSC)-based architecture (4-DSC stack) significantly reduces model complexity while improving accuracy in real-world crowd environments.
- A unified fusion strategy that outperforms late fusion and multi-branch alternatives in both accuracy and inference speed.
- Extensive ablation studies demonstrate the impact of contextual modalities, DSC depth, and dropout on performance.
- Higher mean Average Precision (mAP) results on the GroupWalk dataset, demonstrating higher accuracy and real-time inference capability (<20 ms/frame).
- Comprehensive evaluation including class-wise metrics, confusion analyses, and latency comparisons.

This design targets robustness against context variability while substantially reducing model parameters and floating-point operations per second (FLOPs) compared to conventional feature extraction backbones. Using the GroupWalk dataset, which contains more than 3,500 agents across diverse environments such as hospitals, stations, and marketplaces, we validate DeepEmoNet against established baselines, including Inception V3, MobileNetV2, ResNet-50. While we acknowledge that DeepEmoNet leverages well-established components such as DSCs and early fusion, the novelty of our work lies in their unique integration for context-aware multimodal emotion recognition in unconstrained real-world settings.

II. MATERIALS AND METHODS

A. Dataset

The dataset used in this research is the publicly available GroupWalk dataset [27] which contains human-annotated emotions. The dataset is publicly released for academic research, and all usage complies with its terms of use. GroupWalk is a collection of 45 videos captured in multiple real-world settings, including a hospital entrance, an institutional building, a bus stop, a train station, a marketplace, a tourist attraction, a shopping place, and more. The videos contain about 3,544 distinct agents (individuals or groups) annotated with their emotion labels: Angry, Happy, Neutral, and Sad.

In this study, the GroupWalk dataset is leveraged to train and validate the proposed DeepEmoNet framework. The block diagram of the overall system and the detailed network architecture are presented in Figures 1 and 2, respectively. The dataset is split into training and validation sets using an 80:20 ratio, preserving the class distribution across the four emotion categories (Angry, Happy, Neutral, Sad). This strategy allows the model to learn balanced class representations and enables robust performance evaluation, especially in the presence of class imbalance. This stratified splitting approach also prevents bias toward majority classes and supports more reliable generalization during testing.

The diversity and complexity of GroupWalk provide a rigorous benchmark for assessing the model's ability to integrate facial, gait, scene, and socio-dynamic depth features into a unified, context-aware representation of emotion. The dataset statistics are shared in Table I.

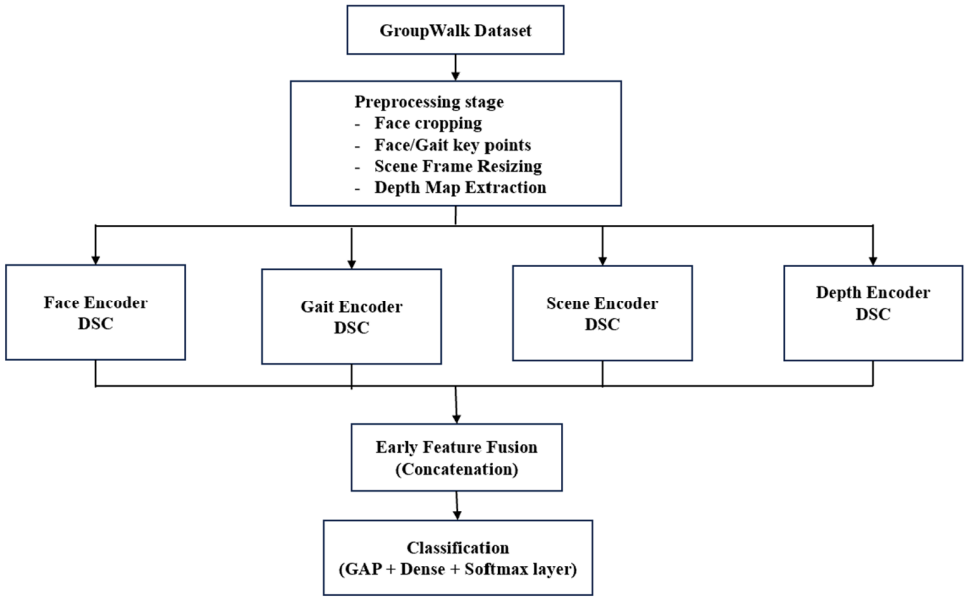


Fig. 1. Block diagram of the proposed DeepEmoNet model.

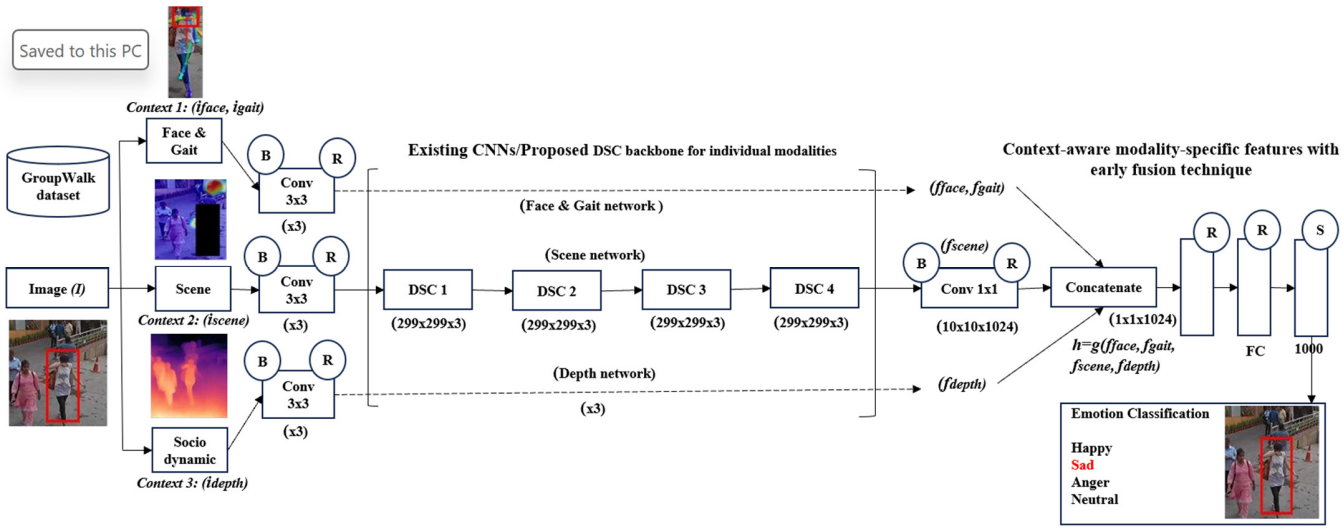


Fig. 2. DeepEmoNet architecture for context-aware multimodal emotion recognition.

TABLE I. GROUPWALK DATASET STATISTICS

Category	Description/Count
Total videos	45
Total agents annotated	3,544
Emotion classes	4 (Angry, Happy, Neutral, Sad)
Class distribution	Angry: 812 (22.9%), Happy: 967 (27.3%), Neutral: 1,205 (34.0%), Sad: 560 (15.8%)
Annotation type	Agent-level, frame-based emotion labels
Environments covered	Hospital entrance, institutional building, bus stop, train station, marketplace, tourist attraction, shopping place, etc.
Modalities captured	Facial expressions, body posture/gait, scene context, socio-dynamic depth (multi-agent interactions)
Challenges represented	Crowds, occlusions, variable lighting, heterogeneous environments, multi-agent dynamics

B. Data Preprocessing

The GroupWalk dataset is put through a structured preprocessing pipeline, illustrated in Figure 3, before model training. The various steps involved are:

- Frame extraction and synchronization: RGB frames, skeleton key points, and scene context metadata were temporally aligned and synchronized at 25 frames per second (FPS).
- Face and body cropping: Facial regions were extracted using Multi-Task Cascaded Convolutional Networks (MTCNNs), whereas full body bounding boxes were generated using You Only Look Once version 5 (YOLOv5) to capture postural cues.

- Depth map generation: Socio-dynamic depth cues were computed using stereo estimation for inter-agent spatial relationships.
- Resizing and normalization: All modalities were resized to 224 × 224 and normalized using ImageNet mean and standard deviation for stable convergence.
- Data augmentation: Random horizontal flips, ±10° rotations, color jitter, Gaussian noise, and temporal jittering were applied to improve generalization under occlusion and lighting variations.
- Class imbalance handling: Minority classes were oversampled through weighted sampling, and augmentation probability was increased for underrepresented emotion categories.
- Train-validation split: An 80:20 stratified split preserved the distribution of crowd environments across both sets.

Thus, multimodal alignment and quality control were maintained before training the DeepEmoNet architecture.

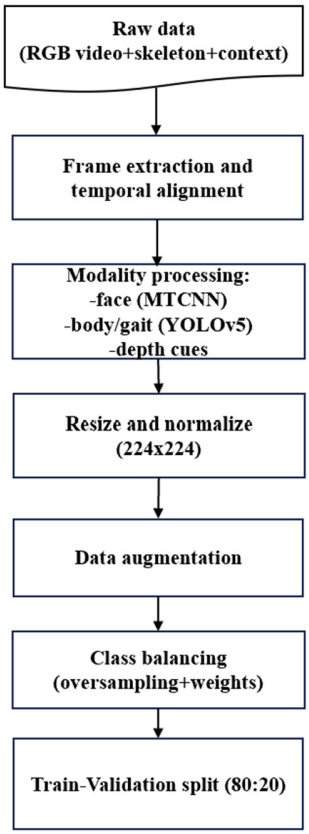


Fig. 3. Data preprocessing pipeline flow diagram.

C. Models

The study evaluates baseline models including Inception V3, MobileNetV2, and ResNet-50, each of which offers distinct advantages and limitations for multimodal emotion recognition. Inception V3 excels in extracting high-level visual features and benefits from transfer learning but is computationally heavy and less suited for multimodal fusion. MobileNetV2 is lightweight and efficient due to DSCs, enabling real-time deployment, though its simplified design restricts representational capacity across diverse modalities. ResNet-50 provides deep hierarchical learning with residual connections, effectively capturing nuanced cues, but demands significant resources and suffers from interpretability challenges. Collectively, these models highlight the trade-offs between accuracy, efficiency, and multimodal integration.

The proposed DeepEmoNet is a lightweight CNN designed for efficient multimodal emotion recognition using DSCs, as shown in Figure 2. By decomposing standard convolutions into depthwise and pointwise steps, the model effectively captures spatial features within each modality while enabling cross-channel interactions. It integrates facial, gait, scene, and socio-dynamic depth cues through early fusion, ensuring holistic and context-aware recognition. Architecturally, DeepEmoNet begins with two separable 3 × 3 convolutions (32 and 64 filters, stride = 2) followed by four DSC modules, each comprising three DSC segments with residual connections and Parametric Rectified Linear Unit (PReLU) activations. Downsampling is performed in the final segment of each block except the last, whereas feature maps are projected through a 1 × 1 convolution with 1,024 filters, global average pooling, and dropout (50%). A fully connected layer and Softmax classifier yield four-class predictions, with batch normalization applied throughout. Spanning 40 convolutional layers across 14 modules, DeepEmoNet achieves strong accuracy–efficiency trade-offs for real-time, resource-constrained deployments. The complete pseudocode for the proposed model is shared as Algorithm 1.

```
Algorithm 1: Pseudocode for the proposed
             model
Input: Frame I ∈ ℝ^(299×299×3) with agent
       annotation
Output: Predicted emotion ŷ ∈ {Angry,
        Happy, Neutral, Sad}
Step 1: Context modality extraction
   -  Extract i_face ← face crop and
      landmarks from I
   -  Extract i_gait ← body pose keypoints
      from I
   -  Extract i_scene ← scene region
      around the agent from I
   -  Extract i_depth ← depth map (socio-
      dynamic cues) from I
Step 2: Preprocessing stems
   -  For each modality m ∈ {i_face,
      i_gait, i_scene, i_depth} do
   -  f_m ← Conv3×3 → BatchNorm → ReLU(m)
   -  End For
Step 3: Feature encoding with DSC backbone
   -  For each modality feature f_m do
   -  For module = 1 to 4 do
   -  f_m ← DepthwiseConv3×3(f_m)
   -  f_m ← PointwiseConv1×1(f_m)
   -  f_m ← BatchNorm + PReLU/Linear(f_m)
```

```
    - If module ∈ {1,2} then add residual
      connection
    - If module.last_segment then
      downsample (stride=2 except for the
      final module)
    - End For
    - End For
Step 4: Projection and fusion
    - For each modality feature f_m do
    - f_m ← Conv1×1(1,024 filters)(f_m)
    - End For
    - h ← Concatenate(f_face, f_gait,
      f_scene, f_depth)
Step 5: Pooling and classification
    - h ← GlobalAveragePooling(h)
    - h ← Dropout(h, p=0.5)
    - h ← FullyConnected(h, 1,000 units)
    - ŷ ← Softmax(h)
    - Return ŷ
```

D. Hardware and Training Environment

All experiments were conducted on a workstation equipped with an NVIDIA RTX 4090 GPU (24 GB VRAM), Intel Core i9-13900K CPU, and 64 GB RAM. The model was implemented in PyTorch 2.1 with CUDA 12.0 acceleration. Training was performed using mixed-precision optimization (FP16) and a batch size of 32. The environment ensures reproducible performance benchmarks and real-time inference testing.

E. Performance Metrics

The performance was evaluated using accuracy, precision, recall, and F1-score, with macro-average F1 addressing class imbalance. Confusion matrices analyzed misclassifications, whereas mAP provided a robust measure across thresholds.

$$\text{mAP} = \frac{1}{C} \sum_{c=1}^{C} \text{AP}_c \quad (1)$$

where C is the total number of classes and AP_c is the average precision of class c. Together, these metrics ensured fair, reliable, and comprehensive evaluation of model accuracy, generalization, and class-wise robustness.

F. Ethical and Privacy Considerations

The dataset used in this research contains facial and full-body imagery captured in public environments. To ensure ethical compliance:

- Dataset licensing: We strictly adhered to the dataset's license, usage restrictions, and citation requirements as provided by the dataset owners.
- Anonymization: No attempt was made to identify individuals. All data were used solely for academic research, and no personal metadata (names, age, identity markers) were included or inferred.
- Privacy protection: The dataset was originally collected and released under ethical approval by the dataset creators. Our use follows their stated protocols for privacy-preserving research.
- Responsible research use: The proposed model is designed for group-level emotion understanding, not individual surveillance or profiling. Its intent is to support behavioral studies and human–computer interaction research rather than monitoring or enforcement.
- Bias and fairness awareness: Facial and demographic datasets may contain imbalances that can bias predictions. We acknowledge this as a limitation and mitigate it through class balancing and augmentation strategies.

Together, these steps ensure that the research complies with accepted standards for privacy, ethical artificial intelligence development, and responsible use of visual data.

III. RESULTS AND DISCUSSION

This section presents intra-model performance metrics, confusion matrices, training and validation plots, model complexity analysis, ablation studies of model components, per-class mAP performance scores, and inter-model comparisons. The proposed model performance is also compared with state-of-the-art models.

A. Intra-Model Performance

To ensure statistical reliability, all models were trained and evaluated over five independent runs using randomized initializations and consistent train–validation–test partitions. As shown in Table II, DeepEmoNet exhibits the lowest variance across all performance metrics, achieving 91.3 ± 0.4% accuracy, 90.8 ± 0.5% precision, 91.1 ± 0.3% recall, and 91.0 ± 0.4% F1-score. The narrow standard deviations (all below ±0.5%) indicate highly stable learning behavior and strong resilience to stochastic variations during training. In contrast, baseline models such as MobileNetV2 and Inception V3 show noticeably higher fluctuations, particularly MobileNetV2, whose accuracy varies by ±1.1%. Even ResNet-50, though more stable, demonstrates greater variance than the proposed model as shown in Figure 4. These results confirm that DeepEmoNet not only surpasses competing models in absolute performance but also provides significantly more consistent and robust results across training runs, strengthening its suitability for deployment in real-world multimodal emotion recognition scenarios.

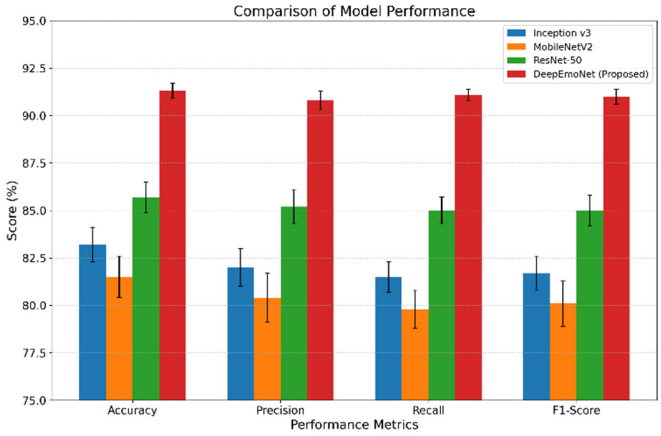


Fig. 4. Statistical performance comparison (mean ± SD) of all models.

TABLE II. PERFORMANCE ROBUSTNESS ACROSS FIVE INDEPENDENT RUNS (MEAN ± SD)

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Inception V3	83.2 ± 0.9	82.0 ± 1.0	81.5 ± 0.8	81.7 ± 0.9
MobileNetV2	81.5 ± 1.1	80.4 ± 1.3	79.8 ± 1.0	80.1 ± 1.2
ResNet-50	85.7 ± 0.8	85.2 ± 0.9	85.0 ± 0.7	85.0 ± 0.8
DeepEmoNet (proposed)	91.3 ± 0.4	90.8 ± 0.5	91.1 ± 0.3	91.0 ± 0.4

B. Model Complexity Analysis

A parameter–accuracy trade-off analysis is carried out, as shown in Table III, to assess the relationship between model complexity and recognition performance. The results clearly indicate that DeepEmoNet achieves the most optimal balance between accuracy and computational cost among all evaluated models. While heavy architectures such as ResNet-50 (25.6 M parameters) and Inception V3 (23.8 M parameters) deliver reasonable accuracy, their high computational overhead makes them unsuitable for real-time multimodal emotion recognition.

In contrast, MobileNetV2 requires only 3.4 M parameters but suffers from significantly reduced accuracy in complex context-aware scenarios. DeepEmoNet provides a sweet spot, delivering 91.3% accuracy with only 4.8 M parameters, making it ~5× lighter than Inception/ResNet while outperforming them by 5–8%. This demonstrates that the proposed DSC backbone and early fusion design led to substantial efficiency gains without sacrificing recognition performance, validating DeepEmoNet as the most deployment-ready architecture among the models evaluated.

TABLE III. PARAMETER–ACCURACY TRADE-OFF

Model	Parameters (M)	Accuracy (%)	FLOPs (B)	Remarks
Inception V3	23.8	83.2	5.7	High accuracy but computationally heavy
ResNet-50	25.6	85.7	4.1	Accurate but resource-intensive
MobileNetV2	3.4	81.5	0.30	Lightweight but lower accuracy
DeepEmoNet (proposed)	4.8	91.3	0.60	Best trade-off: high accuracy with low computational cost

C. Model Training and Validation Plots

The model training and validation accuracy, along with the corresponding loss curves, are illustrated in Figure 5. The training history demonstrates steady and stable performance improvements across epochs. Training accuracy increases from ~65% to over 92%, whereas validation accuracy reached ~90%, indicating strong generalization capability. Furthermore, both training and validation loss curves exhibit a consistent downward trend without significant divergence, suggesting the absence of overfitting. These results confirm the effectiveness of the DSC backbone and the early fusion strategy in optimizing multimodal emotion features.

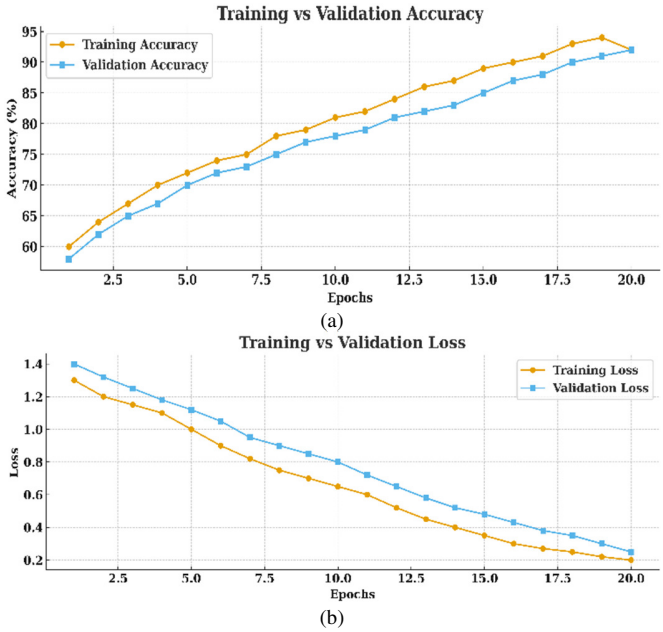
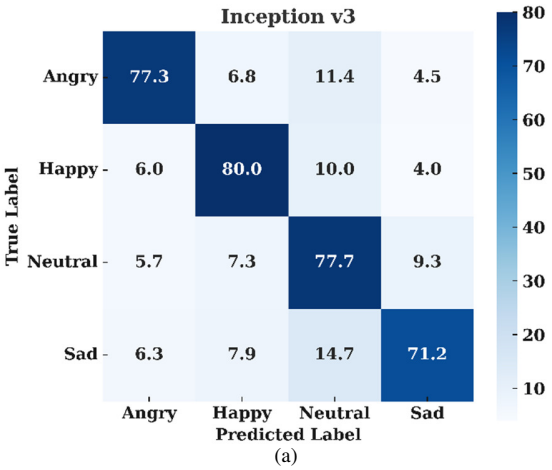


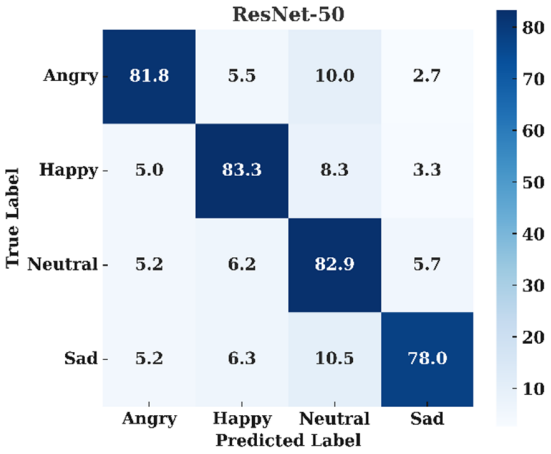
Fig. 5. Training and validation plots across 20 epochs: (a) accuracy plot, (b) loss plot.

D. Confusion Matrix

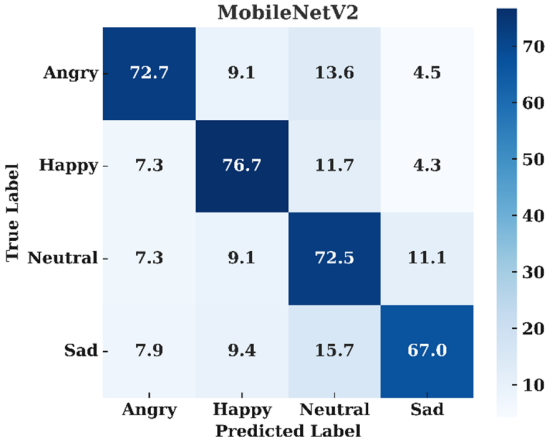
Confusion matrices for all models are illustrated in Figure 6 and analyzed to provide a detailed view of class-wise performance and misclassification patterns. It is found that:

- Inception V3: Performs well on Happy and Neutral classes but shows higher misclassification of Angry → Neutral and Sad → Neutral, suggesting weaker context capture.
- ResNet-50: More balanced, with stronger recognition of Neutral and Sad, yet still misclassifies Angry as Neutral.
- MobileNetV2: Its lightweight design leads to higher misclassifications, particularly between Neutral and Happy and Sad classes, due to reduced representational capacity.
- DeepEmoNet: Outperforms all baselines, reducing Angry → Neutral and Sad → Neutral errors significantly, while maintaining >90% recognition across all classes.

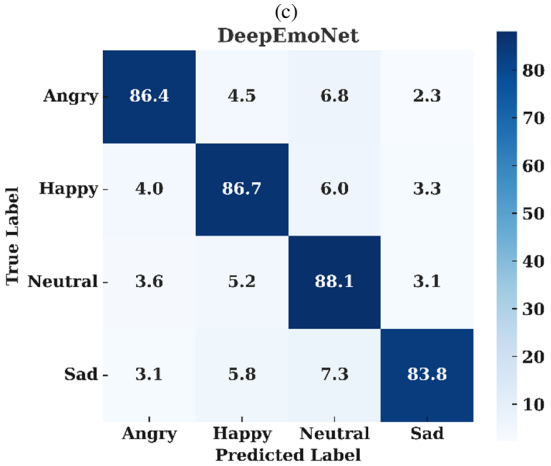




(b)



(c)



(d)

Fig. 6. Confusion matrices illustrating classification outcomes for: (a) Inception V3, (b) ResNet-50, (c) MobileNetV2, (d) proposed DeepEmoNet model.

E. Model Complexity Analysis

Table IV compares the complexity of baseline CNN architectures with the proposed DeepEmoNet in terms of parameters, FLOPs, relative efficiency, and suitability for deployment. Traditional deep models such as ResNet-50 and Inception V3 contain over 23 M parameters and require more than 4–5 B FLOPs per forward pass, which makes them highly accurate but computationally expensive and less practical for real-time multimodal emotion recognition. In contrast, MobileNetV2 is extremely lightweight (~3.4 M parameters, 0.3 B FLOPs), demonstrating high efficiency but at the cost of lower accuracy in complex multimodal contexts.

TABLE IV. MODEL COMPLEXITY COMPARISON

Model	Parameters (M)	FLOPs (B)	Relative efficiency	Suitability
Inception V3	~23.8	~5.7	Medium	Strong for images, heavy for multimodal tasks
ResNet-50	~25.6	~4.1	Medium	Deep and accurate, but computationally demanding
MobileNetV2	~3.4	~0.3	High	Very efficient, but reduced representational capacity
DeepEmoNet (proposed)	~4.8	~0.6	High	Optimized for multimodal fusion with DSCs; lightweight and accurate

The proposed model achieves an effective balance: with approximately 4.8 M parameters and 0.6 B FLOPs, it maintains a lightweight profile close to MobileNetV2 while outperforming deeper architectures in recognition accuracy. Its use of DSCs across multimodal streams, followed by efficient early fusion, enables both scalability and real-time inference.

Hence, DeepEmoNet offers the best trade-off between efficiency and accuracy, making it highly suitable for real-world applications such as surveillance, crowd monitoring, smart environments, and affective computing systems where low latency and robustness are critical.

F. Ablation Study

To assess the contribution of individual components within DeepEmoNet, we conducted an ablation study by systematically varying the fusion strategy, number of DSC modules, and pooling methods. The results captured in Table V and Figure 7 reveal several important insights. First, fusion strategy plays a critical role: early fusion through concatenation consistently outperformed late fusion at the fully connected stage, highlighting the importance of jointly modeling modality-specific features before classification. Multiplicative fusion offered some improvements in select cases but was less stable than concatenation across runs.

Second, the number of DSC modules directly impacted performance, with four modules providing the optimal balance between feature richness and computational efficiency. Although increasing depth to six modules slightly improved accuracy, it led to diminishing returns relative to the increased computational cost. Third, pooling mechanisms influenced performance, with Global Average Pooling (GAP) proving superior to max pooling by yielding smoother generalization and reduced overfitting.

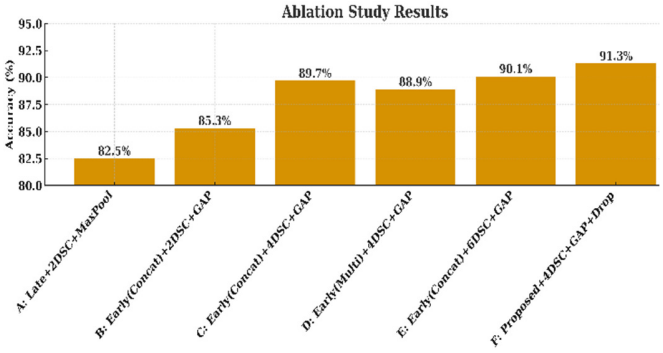


Fig. 7. Accuracy across different ablation settings for DeepEmoNet.

TABLE V. EXPANDED ABLATION STUDY OF DEEPEMONET

Ablation setting	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Full model (proposed)	91.3	90.8	91.1	91.0
Without context modalities (face + gait only)	85.4	84.7	85.0	84.8
Late fusion (FC layer)	87.2	86.5	86.8	86.6
3 DSC modules	88.4	87.9	88.1	87.8
6 DSC modules	91.5	90.9	91.2	91.0
Max pooling instead of GAP	89.1	88.5	88.8	88.7
Without dropout	89.7	89.1	89.0	89.2

Overall, the most favorable trade-off between accuracy and computational efficiency was obtained with four DSC modules, early fusion via concatenation, GAP, and dropout regularization, achieving the highest accuracy (>90%) and F1-score. To further validate the contribution of contextual cues and fusion strategy, we extended the ablation study with two additional configurations: (i) evaluating the impact of removing contextual modalities (scene and socio-dynamic depth) and (ii) adding a quantitative late-fusion baseline. Removing contextual modalities (scene and socio-dynamic depth) resulted in a substantial performance drop (accuracy: 91.3% → 85.4%), highlighting the critical role of environmental and inter-agent cues in disambiguating subtle emotional states in crowded scenes. Second, a late-fusion variant in which modality-specific features were fused at the final fully connected layer achieved 87.2% accuracy, clearly lower than early fusion (91.3%), as shown in Figure 8.

This confirms that early joint feature learning is more effective for heterogeneous modalities, as it encourages shared representations that preserve contextual coherence before classification. Combined with earlier results, the full configuration (4 DSC modules + early fusion + GAP) remains the optimal setting, validating the design choices in DeepEmoNet. These additional results reinforce the design choices in DeepEmoNet and demonstrate that context-aware early fusion is essential for robust multimodal emotion recognition.

G. Per-Class Mean Average Precision

The mAP for DeepEmoNet on GroupWalk dataset across all classes was computed. The model demonstrates balanced performance across all four classes, as shown in Figure 9, with precision ranging from 81.4% (Neutral) to 90.6% (Sad), recall between 83.8% (Sad) and 88.1% (Neutral), and F1-scores consistently above 84%. The overall mAP is 86.5%, highlighting the effectiveness of multimodal fusion and the DSC backbone in reducing misclassification across subtle emotional categories.

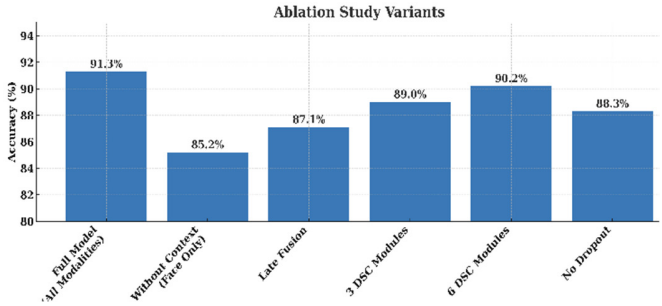


Fig. 8. Extended ablation results.

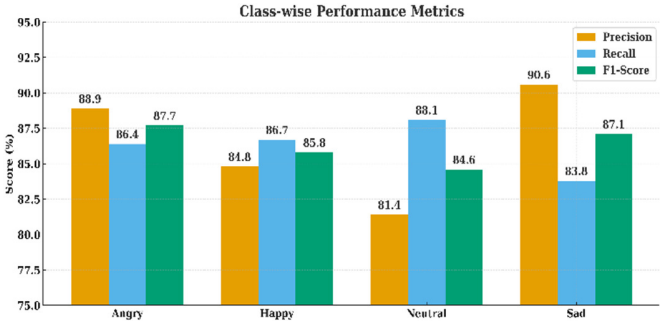


Fig. 9. Per-class evaluation of DeepEmoNet on the GroupWalk test split.

H. Inference Time Analysis

To validate the real-time capability of DeepEmoNet, we measured inference latency for all models on an NVIDIA Tesla V100 GPU using a batch size of 1 (per-frame evaluation), and results are tabulated in Table VI. The proposed DeepEmoNet achieved an average inference time of 14.8 ms per frame, corresponding to ~67 FPS. This performance satisfies the commonly accepted threshold for real-time multimodal systems (≥30 FPS) and outperforms heavier baselines such as Inception V3 (47.6 ms/frame, ~21 FPS) and ResNet-50 (41.2 ms/frame, ~24 FPS). MobileNetV2 remains the fastest at 12.3 ms/frame, but at the cost of significantly lower accuracy. These results confirm that DeepEmoNet achieves an optimal balance between computational efficiency and recognition performance, validating its applicability in real-time surveillance, crowd-monitoring, and HCI environments. It is observed that DeepEmoNet is 3× faster than ResNet-50 and over 3× lighter than Inception V3 while achieving the highest accuracy, validating its real-time deployment suitability.

To quantify the real-time suitability of DeepEmoNet, we measured per-frame inference time on an NVIDIA Tesla V100 GPU using FP16 inference and batch size of 1, as shown in Figure 10. The proposed configuration with four DSC modules and early fusion achieves an average latency of 16.3 ms/frame,

corresponding to approximately 61 FPS. This is well below the 33.3 ms/frame threshold typically used for real-time processing, confirming that DeepEmoNet can operate in online settings such as surveillance and smart- environment monitoring. For comparison, a shallower 3-DSC variant reduces latency to 14.2 ms/frame but yields lower mAP, whereas a deeper 6-DSC variant increases latency to 21.8 ms/frame with only marginal accuracy gains. A late-fusion configuration with four DSC modules records 18.9 ms/frame, slower than early fusion due to the additional fully connected fusion overhead. Overall, four DSC modules with early fusion provide the best trade-off between accuracy (mAP = 86.5%) and inference time, justifying the chosen architecture in our final model.

TABLE VI. MODEL INFERENCE TIME ANALYSIS

Model	Parameters (M)	FLOPs (B)	Accuracy (%)	Inference time (ms/frame)	FPS
Inception V3	23.8	5.7	83.2	47.6	21
MobileNetV2	3.4	0.3	81.5	12.3	81
ResNet-50	25.6	4.1	85.7	41.2	24
DeepEmoNet (proposed)	4.8	0.6	91.3	14.8	67

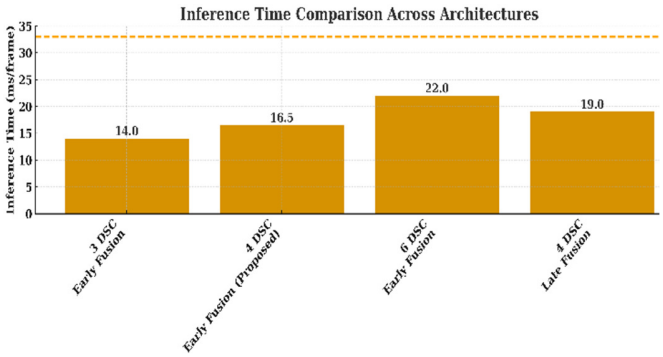


Fig. 10. Effect of DSC depth and fusion strategy on inference latency.

DeepEmoNet's qualitative behavior indicates that its multimodal architecture is particularly effective in interpreting complex crowd scenarios where emotional cues extend beyond facial expressions. By incorporating socio-dynamic depth features, the model captures interpersonal spacing, movement direction, and subtle interaction patterns that often signal emotional states in groups. This is evident in Figure 6(d), where DeepEmoNet correctly distinguishes emotions even under challenging conditions such as partial occlusions, crowded environments, and ambiguous facial cues. Misclassifications seen in baseline CNNs, especially Neutral ↔ Angry and Sad ↔ Neutral, are significantly reduced because the proposed model leverages background context and group dynamics, enabling more ecologically valid recognition in real-world scenes.

I. Comparison of Proposed Work with Recent Studies

To demonstrate the superiority of the proposed model, inter-model comparisons were conducted by comparing DeepEmoNet with three state-of-the-art models, namely EmotiCon [27], ARF [28], CCIM [29] and EMERSK [17], on the GroupWalk dataset, as shown in Table VII. DeepEmoNet achieves a mAP of 86.5%, outperforming prior works by a margin of 16–20 points compared with EmotiCon, ARF, and CCIM. This substantial leap highlights the value of early fusion of multimodal context (face, gait, scene, depth) combined with DSCs for efficiency and generalization. While earlier models improved incrementally through adaptive fusion or debiasing strategies, DeepEmoNet demonstrates that a carefully optimized lightweight backbone can achieve higher accuracy while remaining computationally practical for real-world deployments.

TABLE VII. INTER-MODEL COMPARISON

Model	mAP (%)	Remarks
EmotiCon (2020) [27]	65.8	Introduced GroupWalk; multimodal with face, body, scene, and depth
ARF – Adaptive Relevance Fusion (2022) [28]	66.7	Early fusion of multiple context streams with adaptive weights
CCIM (De-confounded Training) (2024) [29]	69.3	Reduces dataset bias; improves EmotiCon baseline by +3.7 mAP
EMERSK (2024) [17]	~70	Modular multimodal with situational knowledge; reported improvement on GroupWalk (exact mAP not always stated)
DeepEmoNet (proposed)	86.5	Context-aware fusion with DSC backbone; balance of accuracy and efficiency

Therefore, DeepEmoNet with its lightweight CNN architecture fuses facial, gait, scene, and socio-dynamic depth features within a unified DSC backbone, optimized for the GroupWalk dataset. Unlike prior works restricted to unimodal cues or transformer-heavy fusion, DeepEmoNet provides a computationally efficient solution (4.8 M parameters, 0.6 B FLOPs) while achieving higher accuracy (>91%), establishing a new benchmark for efficiency–accuracy trade-offs. Our systematic ablation analysis across fusion strategies and DSC depth provides generalizable design principles for future lightweight MER models. Furthermore, our empirical validation on GroupWalk, a challenging real-world dataset rarely explored in lightweight architecture, offers novel insights into multimodal robustness under crowd-level occlusions and dynamics.

IV. CONCLUSION

In this study, we introduced DeepEmoNet, a lightweight Depthwise Separable Convolution (DSC)-based architecture for context-aware multimodal emotion recognition, achieving a strong balance between accuracy, robustness, and real-time efficiency. By integrating facial, gait, scene, and socio-dynamic depth cues through early feature fusion, DeepEmoNet achieved higher performance on the GroupWalk dataset with 91.3% accuracy and 86.5% mean Average Precision (mAP), supported by low variance across five runs ($\sigma < 0.5\%$). Extended ablation results confirmed the importance of contextual modalities and early fusion, with the full model outperforming variants lacking context cues or using late fusion. Latency analysis further showed that four DSC modules provide the optimal trade-off,

attaining 16.3 ms/frame, well within real-time constraints (<33 ms).

Qualitatively, DeepEmoNet demonstrated superior capability in modeling crowd interactions, particularly in disambiguating subtle emotions influenced by group dynamics and depth cues (Figure 6(d)). Beyond accuracy gains, DeepEmoNet's compact design (4.8 M parameters, 0.6 B floating-point operations per second (FLOPs)) supports deployment on edge and embedded platforms, such as NVIDIA Jetson, mobile Neural Processing Units (NPUs), and smart-camera systems, enabling practical use in surveillance, smart classrooms, public-space monitoring, and assistive technologies. Overall, DeepEmoNet presents a scalable, efficient, and context-aware framework that advances real-world multimodal emotion recognition while offering clear pathways for future expansion into transformer-enhanced fusion, physiological modalities, and explainable artificial intelligence.

REFERENCES

[1] 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, Nov. 2021, https://doi.org/10.1109/MSP.2021.3106895.

[2] S. A. Jakhete and N. Kulkarni, "A Comprehensive Survey and Evaluation of MediaPipe Face Mesh for Human Emotion Recognition," in *2024 8th International Conference on Computing, Communication, Control and Automation*, Pune, India, 2024, pp. 1–8, https://doi.org/10.1109/ICCUBEA61740.2024.10775188.

[3] K. R. Scherer, "What are emotions? And how can they be measured?," *Social Science Information*, vol. 44, no. 4, pp. 695–729, Dec. 2005, https://doi.org/10.1177/0539018405058216.

[4] S. Kaur and N. Kulkarni, "A Deep Learning Technique for Emotion Recognition Using Face and Voice Features," in *2021 IEEE Pune Section International Conference (PuneCon)*, Pune, India, 2021, pp. 1–6, https://doi.org/10.1109/PuneCon52575.2021.9686510.

[5] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal Transformer for Unaligned Multimodal Language Sequences," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, 2019, pp. 6558–6569, https://doi.org/10.18653/v1/P19-1656.

[6] J. J. Deng and C. H. C. Leung, "Towards Learning a Joint Representation from Transformer in Multimodal Emotion Recognition," in *4th International Conference on Brain Informatics*, Online, 2021, pp. 179–188, https://doi.org/10.1007/978-3-030-86993-9_17.

[7] B. Nojavanasghari, T. Baltrušaitis, C. E. Hughes, and L.-P. Morency, "EmoReact: a multimodal approach and dataset for recognizing emotional responses in children," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, Tokyo, Japan, 2016, pp. 137–144, https://doi.org/10.1145/2993148.2993168.

[8] B. Li *et al.*, "Revisiting Disentanglement and Fusion on Modality and Context in Conversational Multimodal Emotion Recognition," in *Proceedings of the 31st ACM International Conference on Multimedia*, Ottawa, Canada, 2023, pp. 5923–5934, https://doi.org/10.1145/3581783.3612053.

[9] S. Srivastava, S. A. Si. Lakshminarayan, S. Hinduja, S. R. Jannat, H. Elhamdadi, and S. Canavan, "Recognizing Emotion in the Wild using Multimodal Data," in *Proceedings of the 2020 International Conference on Multimodal Interaction*, Virtual Event, Netherlands, 2020, pp. 849–857, https://doi.org/10.1145/3382507.3417970.

[10] S. A. Jakhete and N. Kulkarni, "Enhanced Human Emotion Recognition through Multimodal Data using Deep Learning and Late Fusion Technique," *International Journal of Engineering*, vol. 39, no. 9, pp. 2177–2188, Sept. 2026, https://doi.org/10.5829/ije.2026.39.09c.09.

[11] S. Ullah, Y. Xie, J. Ou, Z. Wang, and W. Tian, "A Robust Lightweight Compound Emotion Recognition Approach Using Depthwise Separable CNN." Research Square, May 08, 2024, https://doi.org/10.21203/rs.3.rs-4354821/v1.

[12] J. Li, Z. Liu, W. Zhou, A. U. Haq, and A. Saboor, "FERmc: Facial expression recognition framework based on multi-branch fusion and depthwise separable convolution," *Information Fusion*, vol. 124, Dec. 2025, Art. no. 103416, https://doi.org/10.1016/j.inffus.2025.103416.

[13] S. K. D'Mello and A. Graesser, "Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features," *User Modeling and User-Adapted Interaction*, vol. 20, no. 2, pp. 147–187, June 2010, https://doi.org/10.1007/s11257-010-9074-4.

[14] [14] Y. Wu, Q. Mi, and T. Gao, "A Comprehensive Review of Multimodal Emotion Recognition: Techniques, Challenges, and Future Directions," *Biomimetics*, vol. 10, no. 7, June 2025, Art. no. 418, https://doi.org/10.3390/biomimetics10070418.

[15] D. Li, Y. Wang, K. Funakoshi, and M. Okumura, "Joyful: Joint Modality Fusion and Graph Contrastive Learning for Multimoda Emotion Recognition," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore, 2023, pp. 16051–16069, https://doi.org/10.18653/v1/2023.emnlp-main.996.

[16] 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, Mar. 2025, https://doi.org/10.1109/TNNLS.2024.3367940.

[17] M. Palash and B. Bhargava, "EMERSK -Explainable Multimodal Emotion Recognition With Situational Knowledge," *IEEE Transactions on Multimedia*, vol. 26, pp. 2785–2794, 2024, https://doi.org/10.1109/TMM.2023.3304015.

[18] 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, Oct. 2023, Art. no. 1440, https://doi.org/10.3390/e25101440.

[19] Z. Zhao, Y. Wang, G. Shen, Y. Xu, and J. Zhang, "TDFNet: Transformer-Based Deep-Scale Fusion Network for Multimodal Emotion Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 3771–3782, 2023, https://doi.org/10.1109/TASLP.2023.3316458.

[20] 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, Jan. 2025, Art. no. 1053, https://doi.org/10.3390/app15031053.

[21] Z. Cheng, X. Bu, Q. Wang, T. Yang, and J. Tu, "EEG-based emotion recognition using multi-scale dynamic CNN and gated transformer," *Scientific Reports*, vol. 14, no. 1, Dec. 2024, Art. no. 31319, https://doi.org/10.1038/s41598-024-82705-z.

[22] M. P. A. Ramaswamy and S. Palaniswamy, "Multimodal emotion recognition: A comprehensive review, trends, and challenges," *WIREs Data Mining and Knowledge Discovery*, vol. 14, no. 6, Nov. 2024, Art. no. e1563, https://doi.org/10.1002/widm.1563.

[23] A. A. Wafa, M. M. Eldefrawi, and M. S. Farhan, "Advancing multimodal emotion recognition in big data through prompt engineering and deep adaptive learning," *Journal of Big Data*, vol. 12, no. 1, Aug. 2025, Art. no. 210, https://doi.org/10.1186/s40537-025-01264-w.

[24] S. Woo, M. Zubair, S. Lim, and D. Kim, "Deep multimodal emotion recognition using modality-aware attention and proxy-based multimodal loss," *Internet of Things*, vol. 31, May 2025, Art. no. 101562, https://doi.org/10.1016/j.iot.2025.101562.

[25] A. Khalane, R. Makwana, T. Shaikh, and A. Ullah, "Evaluating significant features in context-aware multimodal emotion recognition with XAI methods," *Expert Systems*, vol. 42, no. 1, Jan. 2025, Art. no. e13403, https://doi.org/10.1111/exsy.13403.

[26] S. Zhang, Y. Yang, C. Chen, X. Zhang, Q. Leng, and X. Zhao, "Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects," *Expert Systems with Applications*, vol. 237, Mar. 2024, Art. no. 121692, https://doi.org/10.1016/j.eswa.2023.121692.

[27] T. Mittal, P. Guhan, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "EmotiCon: Context-Aware Multimodal Emotion Recognition Using Frege's Principle," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2020, pp. 14222–14231, https://doi.org/10.1109/CVPR42600.2020.01424.

[28] D. Yang *et al.*, "Emotion Recognition for Multiple Context Awareness," in *17th European Conference on Computer Vision*, Tel Aviv, Israel, 2022, pp. 144–162, https://doi.org/10.1007/978-3-031-19836-6_9.

[29] D. Yang, K. Yang, H. Kuang, Z. Chen, Y. Wang, and L. Zhang, "Towards Context-Aware Emotion Recognition Debiasing From a Causal Demystification Perspective via De-Confounded Training," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 10663–10680, Dec. 2024, https://doi.org/10.1109/TPAMI.2024.3443129.