Group Emotion Recognition using Multimodal Transformers with Hierarchical Aggregation

Gaelle Laetitia MOUAFO MAPIKOU
laetitiamouafo@gmail.com
New York Institute of Technology

Houwei CAO
hcao02@nyit.edu
New York Institute of Technology

***Abstract*—Group emotion recognition (GER) involves modeling emotional dynamics across multiple individuals using diverse data types. We propose a transformer-based framework that integrates visual, audio, and text information through hierarchical feature aggregation and self-attention mechanisms to capture both intra- and inter-modal dependencies. A dedicated fusion layer further refines the joint representations, while multitask learning facilitates concurrent optimization for recognizing emotions on both individual and group levels. This approach improves contextual understanding and generalization across different group sizes. Evaluations on MELD, IEMOCAP, and VGAF datasets show that our approach outperforms strong baselines such as SCFA and VGAFNet. Ablation studies confirm the significance of hierarchical aggregation, multitask learning, and multimodal fusion. Overall, our framework provides a flexible and effective solution for real-world GER applications.**

***Index Terms*—Group Emotion Recognition, Multimodal Transformer, Attention, Fusion, Hierarchical Aggregation**

I. Introduction

In recent years, affective computing has emerged as a crucial area of research, enabling machines to recognize, interpret, and respond to human emotions. Emotion Recognition (ER) plays a vital role in diverse domains, such as enhancing learning experiences in classrooms, improving human-computer interaction, tailoring content recommendations in multimedia platforms, and analyzing group dynamics in social settings. While significant advancements have been made in recognizing individual emotions through modalities such as facial expressions, speech, and text, the task of Group Emotion Recognition (GER) poses additional complexities .

ER has traditionally focused on identifying the affective states of individuals, relying on cues such as facial expressions, vocal tone, and textual sentiment. However, GER introduces a distinct set of challenges that go beyond the sum of individual emotions . Unlike individual ER, GER must account for complex interpersonal dynamics, contextual dependencies, and the collective emotional state that emerges from the interaction among group members. For example, while individual members of a group may exhibit varied expressions, the group's overall emotion may reflect a shared response to a common stimulus, such as agreement, tension, or excitement during a conversation. These unique challenges demand models capable of capturing both individual signals and their interrelations within a shared context, making GER a more nuanced and socially informed task than its individual counterpart.

GER is the task of identifying and analyzing the collective emotional state of a group of individuals based on various data modalities, such as facial expressions, body language, speech, and contextual cues. Similar to individual ER, GER utilizes techniques from computer vision, natural language processing, and audio analysis. Due to the added complexity of assessing the emotion state of a group, fusion of multimodal signals from multiple users and integration of deep learning models developed for different modalities are more critical and challenging for improving the accuracy and robustness of GER. Challenges remain in dataset acquisition and feature harmonization [5]. Data noise and bias present significant hurdles, as real-world data often contain noises, and datasets may have inherent biases that affect model performance. The computational complexity of advanced models, particularly those that combine multiple modalities, requires significant computational resources, making them less accessible for some applications. Ensuring generalizability remains a challenge, as models need to perform well across different datasets and contexts. Integrating multiple modalities effectively is challenging and requires advanced fusion techniques that preserve essential information.

Group emotions are dynamic and context-dependent, requiring a holistic understanding that integrates multiple modalities effectively. Although advances have been made in unimodal approaches [7], recognizing group emotions in dynamic multimodal contexts remains a significant challenge, mostly because the existing models often fail to capture the complexity and nuance of group interactions [5]. Real-world scenarios require the integration of diverse data streams, including text, audio, and visual cues, that interact in intricate ways to form a holistic representation of group emotions. Furthermore, current state-of-the-art models, such as SCFA [9] and VGAFNet [11], exhibit limitations in scalability and accuracy. SCFA, while leveraging speaker-aware cross-modal fusion, is restricted to datasets with detailed speaker annotations. VGAFNet, optimized for audio-visual data, does not incorporate textual information, resulting in incomplete emotion recognition in scenarios where context from speech or text is crucial.

The primary research problem addressed in this paper is the lack of a robust, scalable and accurate framework for GER that integrates text, audio and visual modalities. Existing approaches often prioritize one or two modalities, neglecting the synergistic potential of combining all three. Furthermore, current methods struggle with generalization across diverse datasets and scenarios, limiting their applicability in real-world settings. Our framework incorporates self-attention mecha-

nisms to effectively process and integrate text, audio, and visual features. Leveraging the capabilities of transformers [6], our framework aims to capture rich inter-modal dependencies and deliver good performance. Our methodology involves designing a multimodal architecture that processes each modality using specialized pre-processing pipelines and feeds them into a unified transformer-based fusion model. The framework is evaluated against baselines on MELD, IEMOCAP, and VGAF benchmark datasets. The accuracy and F1 scores are analyzed to validate the effectiveness of our framework. The contributions of this work are summarized as follows:

- A transformer-based architecture for GER that fuses modality-specific features to improve accuracy and robustness;
- A unified preprocessing pipeline for text tokenization, audio feature extraction, and visual data processing, ensuring compatibility across diverse datasets;
- Strong performance on benchmark datasets (MELD, IEMOCAP, VGAF), comparable to or surpassing state-of-the-art methods.

II. Related Work

ER has been extensively studied across multiple modalities, each leveraging different computational techniques to enhance accuracy and robustness. In facial ER, Convolutional Neural Networks (CNNs) have been widely used for valence and arousal estimation [3], [4], [20]. Speech-based methods rely on Recurrent Neural Networks (RNNs) [1] and attention mechanisms to extract temporal emotional cues from voice features, highlighting the importance of prosody and tone in sentiment analysis [12]. Text-based approaches utilize Support Vector Machines (SVMs) , lexicon-based techniques, and advanced BERT transformer models [13], which incorporate contextual and speaker-aware information for more nuanced emotion detection in textual data. Multimodal approaches have recently gained traction in GER to improve the ability to analyze collective emotional states [2], [14]. The effectiveness of GER depends on robust fusion strategies (early,late and hierarchical) to capture intricate cross-modal interactions. Deep learning models, particularly CNNs for visual feature extraction [17], [18], RNNs for sequential data processing [19], and transformers for long-range dependencies [9], [11], play a crucial role in ER. Attention mechanisms further enhance performance by dynamically weighing features across different modalities, ensuring that the most relevant information contributes to emotion classification.

In summary, group emotion recognition has evolved significantly, moving from focusing solely on facial features to incorporating diverse data sources like local object features, scene features, and audio. This shift towards multimodal approaches has improved its accuracy, robustness, and interpretability. Researchers have demonstrated the effectiveness of multimodal fusion methods in achieving a more comprehensive understanding of group emotions by leveraging complementary information from multiple data sources. However, challenges remain, including the difficulty of obtaining and annotating multimodal datasets and harmonizing features across modalities. Despite these hurdles, multimodal fusion is expected to dominate GER research, with future efforts aiming to refine integration techniques and expand datasets for broader applicability.

III. Methodology

A. Problem definition

GER aims to predict the collective emotional state of a group based on individual emotions and their interactions. Unlike individual ER, GER must capture complex dependencies among group members, considering factors such as shared context, social dynamics, and multimodal cues. A unimodal approach, relying only on facial expressions or speech, often fails in real-world scenarios. For example, a group in a meeting might have neutral facial expressions but display excitement through their tone and gestures. Similarly, textual conversations in a chat may lack emotional clarity without vocal intonation or facial expressions. A multimodal approach, which combines visual, audio, and textual data, ensures a more comprehensive understanding of group emotions by leveraging complementary cues, making it robust in diverse settings like meetings, social gatherings.

B. Proposed Framework

We propose a framework built on three mechanisms: Transformers, Attention, and Multitask Learning (see Fig.1). The framework integrates state-of-the-art multimodal transformers for robust group emotion recognition, leveraging advancements in text, audio, and visual modality modeling. Attention mechanisms are further used to enhance the framework's capability by capturing both intramodal and intermodal dependencies. Self-attention mechanisms focus on relationships within a single modality, ensuring comprehensive feature extraction, while cross-modal attention captures interdependencies between modalities, enabling a richer and more holistic understanding of the data. The framework employs multitask learning to improve robustness. By combining GER (primary task), with individual sentiment analysis (auxiliary task), the model benefits from shared learning signals, which improve its predictive accuracy and generalization across related tasks. Together, these components create a powerful system for multimodal emotion analysis.

Our framework addresses Group Emotion Recognition (GER) using a unified multimodal transformer architecture with hierarchical aggregation. It begins by extracting modality-specific features (text, audio, vision) for each individual in a group. These features are encoded using respective encoders (e.g., BERT for text, CNNs for visual and acoustic signals). To handle variable group sizes, we introduce a learned separator token between each individual's representation, ensuring the model distinguishes between different members.

The fusion layer applies intra-modal self-attention to refine representations within each modality, followed by inter-modal self-attention to capture dependencies across modalities. These fused multimodal representations are hierarchically aggregated

across individuals using attention-based pooling, enabling the model to understand shared emotional context and interpersonal dynamics. The prediction layer uses parallel task-specific heads: one for individual emotion recognition (ER) and one for GER. Each individual's prediction is computed independently using their own fused features. For GER, a group-level representation is constructed from the aggregated features of all individuals in the scene.

To train the model, we adopt a multitask learning (MTL) objective combining both individual and group emotion loss. This allows shared learning between tasks, improving both levels of prediction. Our framework supports datasets with variable group sizes and missing modalities, making it robust and scalable.

1) Input Preprocessing Layer: The Input Preprocessing Layer extracts meaningful features from each modality. For the text modality, the BERT language transformer is used to generate contextualized embeddings from dialogue transcripts, capturing both syntactic and semantic nuances. The audio data are converted into spectrograms and processed using the Audio Spectrogram Transformer (AST), which extracts features from the temporal and frequency domains. For visual data, the Vision Transformer (ViT) encodes inputs such as facial expressions and gestures, incorporating spatial and temporal dynamics to capture non-verbal cues effectively.

2) Fusion Layer: The Fusion Layer employs a multi-modal transformer (MMT) to align and integrate representations of the three modalities. This layer uses self-attention and cross-modal attention mechanisms to model intra-/intermodal interactions. By dynamically aligning features across modalities, the Fusion Layer creates a unified representation that captures complex relationships and improves the interpretability of the data.

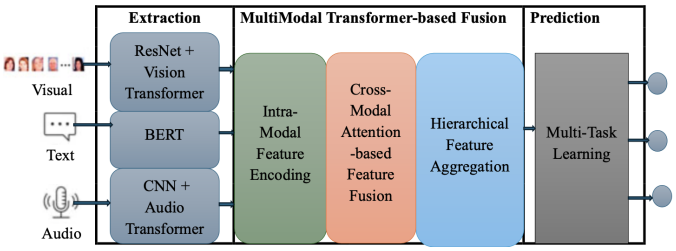

Fig. 1. Architecture of Proposed Multimodal GER Model

Intra-modal Feature Encoding: Within each modality, the extracted features are processed independently through transformer encoders where self-attention mechanisms capture intramodal dependencies by refining the individual modality representations. For modality $M \in \{text, audio, visual\}$

$$M_{enc} = TransformerEncoder(M_{emb}), \tag{1}$$

where M_{emb} represents the input embeddings of the modality, and $M_{enc} \in \mathbb{R}^{n_m \times d_m}$ represents encoded features. A *transformer encoder* consists of a Multi-Head Self-Attention (MHSA) mechanism:

$$MHSA(X) = Concat(head_1, ..., head_h)W_o, \tag{2}$$

$$head_i = Softmax(\frac{Q_i K_i^T}{\sqrt{d_i}})V_i, \tag{3}$$

where $Q_i = XW_i^Q$, $K_i = XW_i^K$, $V_i = XW_i^V$, respectively representing the Query, Key, and Value matrices, and d_i is the dimension of each head. W_i typically represents the learnable weight matrices used in linear transformations for an input sequence X; and a Feed-Forward neural Network (FFN):

$$FFN(X) = ReLU(XW_1 + b_1)W_2 + b_2, \tag{4}$$

where W_1 and W_2 are learnable weight matrices for the first and second linear transformations, respectively; b_1 and b_2 are bias terms.

Cross-Modal Attention-Based Fusion: the encoded features from all three modalities are combined using a multimodal transformer, where cross-attention layers facilitate intermodal interactions. This mechanism allows the model to learn complementary relationships between modalities, improving the shared representation. The attention mechanism assigns higher weights to more informative features, ensuring that critical emotional cues are emphasized in the fusion process. Let $M_{fusion} = [T_{enc}, A_{enc}, V_{enc}]$ be the concatenated features.

$$M_{fused} = TransformerEncoder(M_{fusion}). \tag{5}$$

Hierarchical Aggregation and Final Representation: The refined multimodal representations are progressively aggregated using a hierarchical structure , where lower-level fused features contribute to higher-level contextual embeddings. A fully connected feedforward network (FFN) further refines this aggregated representation before final classification. To transition from individual to group emotion representation, a hierarchical attention mechanism aggregates features from all group members. The model assigns importance weights to each member's contribution based on context and group dynamics, ensuring that dominant emotional signals drive the final prediction. The hierarchical fusion strategy ensures that the most relevant multimodal dependencies are preserved, leading to robust GER.

3) Prediction Layer: The prediction layer of the proposed framework employs a multitask learning (MTL) strategy to enhance group emotion recognition by simultaneously learning individual and group-level emotion representations. This design ensures that the model effectively captures fine-grained emotional cues from individuals while leveraging these insights to infer collective group emotions. The multitask learning mechanism operates through the following steps:

Parallel Task-Specific Heads: the output of the hierarchical fusion layer is fed into two separate branches in the prediction layer: (1) Individual Emotion Prediction Head: A fully connected layer with a softmax activation function is used to classify each individual's emotional state (e.g., happy, sad, angry, etc.). Our framework incorporates individual features

and relevant group context for individual emotion prediction, thus ensuring that the model learns precise individual-level emotion representations; (2) Group Emotion Prediction Head: another fully connected layer processes the fused multimodal representation to predict the overall group emotion. This head considers aggregated intra-group relationships, modeling how individual emotions influence the collective emotional state.

Shared Feature Learning: a shared backbone, consisting of transformer-based fusion layers, learns common features beneficial for both tasks. By jointly optimizing both heads, the model benefits from shared representations, allowing it to generalize better across varying group compositions and contextual emotional expressions.

$$z = Pooling(M_{fused}), \tag{6}$$

where Pooling can be mean pooling or max pooling to summarize the sequence features.

$$\Upsilon_{pred} = Softmax(zW_c + b_c), \tag{7}$$

where W_c and b_c are learnable parameters of the classification layer and Υ_{pred} represents the predicted probabilities for each emotion class. The predicted label and loss function are:

$$\hat{y} = \underset{c}{argmax}(\Upsilon_{pred,c}) \tag{8}$$

$$\mathcal{L} = -\frac{1}{N}\Sigma_{i=1}^{N}\Sigma_{c=1}^{C} y_{i,c} log(\Upsilon_{pred,i,c}), \tag{9}$$

where N is the number of samples, C is the number of classes, $y_{i,c}$ is the ground truth label and $\Upsilon_{pred,i,c}$ is the predicted probability for sample i in class c.

Loss Function Optimization: the multitask learning setup optimizes a combined loss function to balance both individual and group emotion prediction tasks:

$$\mathcal{L}_{Total} = \lambda_1 \mathcal{L}_{individual} + \lambda_2 \mathcal{L}_{group}, \tag{10}$$

where $\mathcal{L}_{individual}$, $\mathcal{L}_{group}$ are the categorical cross-entropy loss for individual/group-level emotion classification, λ_1, λ_2 are weighting factors to control each task's contribution.

IV. Evaluation

A. Datasets

This work considers three publicly available state-of-the-art datasets, namely, IEMOCAP [8], MELD [10] and VGAF [11]. IEMOCAP (Interactive Emotional Dyadic Motion Capture) is a multimodal dataset featuring 151 dyadic conversations with 7,433 utterances, each labeled with one of six emotions: `happy`, `sad`, `angry`, `excited`, `frustrated`, or `neutral`. It also includes annotations for valence, arousal, and dominance. It offers multimodal data: audio, video and motion capture (3D movements of the head, hands, and face). MELD (Multimodal EmotionLines Dataset) is a multimodal conversational dataset that includes audio, visual, alongside with text transcriptions. It features 1,433 dialogues and 13,804 utterances from the TV series "Friends", with multiple speakers involved. Each utterance in the dialogue is labeled with one of seven emotions: `Anger`, `Disgust`, `Sadness`, `Joy`, `Neutral`, `Surprise`, `Fear`. The VGAF (Video Group AFfect) dataset is a multimodal collection comprising 326 original video clips downloaded from youtube, for a total number of 4,183 samples with 587,364 frames. Each video lasts for five seconds on average and is annotated with one of three group-level affective states: `Negative`, `Neutral`, `Positive`. Table I summarizes the caracteristics of these datasets.

TABLE I
DATASET SUMMARY

Dataset	MELD	IEMOCAP	VGAF
Desc	Transcripts	Dyadic	Environment
Instances	13804 utterances	7433 utterances	4183 samples
Modalities	Text/Audio/Visual	Text/Audio/Visual	Audio/Visual
ER Classes	7 classes	6 classes	3 classes
ER Level	Individual	Individual	Group
Group Size	$\approx$ 3 individuals	$\approx$ 2 individuals	$\approx$ 4 individuals

B. Implementation Details

The implementation of the proposed multimodal framework for group emotion recognition is carried out in Python. The framework relies on PyTorch for building and training deep learning models, Hugging Face's Transformers for text tokenization and embedding extraction, Librosa for audio feature extraction, and OpenCV for visual data preprocessing. The framework trains using the Adam optimizer with a learning rate of 0.0001, and a batch size of 32. Training spans 50 epochs, allowing the model to achieve optimal convergence, as evidenced by a steady reduction in the loss function and stabilization of evaluation metrics. Cross-validation is performed using a 5-fold strategy to ensure the robustness and generalizability across diverse datasets.

C. Baselines

SCFA [9] is a model designed for multimodal emotion recognition, focusing on integrating audio, visual, and text data with speaker-aware cross-modal attention. It enhances feature alignment by considering speaker dynamics. VGAFNet [11] specializes in audio-visual group emotion recognition by leveraging visual cues to refine audio representations. We also compare our method with HiMER [1], a recent state-of-the-art model for emotion recognition. HiMER models hierarchical relationships using structured graphs to capture speaker and group-level emotion dependencies in multi-party dialogues. While highly effective in dialogue-centric datasets like IEMOCAP, HiMER lacks the modular flexibility and fusion adaptability offered by our multimodal transformer-based framework, especially in group scenes with variable compositions.

D. Ablation Study

The purpose of this ablation study is to evaluate the contribution of each modality and the specific architectural components to the performance of the proposed framework for GER.

1) Modality Importance: This study aims to analyze the significance of individual modalities and their combinations and compare the framework's performance against baselines . Table II presents the results. In analyzing the importance of each modality across the datasets, distinct insights emerge. For MELD, text proves to be the most critical modality as it encapsulates the primary semantic information within conversational exchanges, enabling a deeper understanding of context and intent. In contrast, IEMOCAP emphasizes the importance of audio, which provides significant emotional cues through tone, pitch, and rhythm, which is particularly relevant in its focus on dyadic interactions. Meanwhile, VGAF highlights the dominance of visual data, which effectively captures group-level nonverbal dynamics such as facial expressions and gestures, making it the most influential modality for understanding collective emotions.

TABLE II
MODALITY IMPORTANCE: TEXT (T), AUDIO (A), VISUAL (V)

	IEMOCAP		MELD		VGAF	
Criteria	**Acc**	**F1**	**Acc**	**F1**	**Acc**	**F1**
T	63.82	62.89	59.32	57.76	-	-
A	49.24	48.66	47.37	44.18	49.11	48.57
V	40.12	38,23	50.37	46.16	70.96	69.98
TA	67.99	66.72	64.97	63.89	-	-
TV	66.81	65.78	61.34	59.74	-	-
AV	57.11	55.23	48.33	44.11	72.56	71.47
TAV	**70.03**	**68.41**	**68.92**	**67.60**	-	-

2) Architectural Components: This study aims to assess the impact of the self-attention mechanism and transformer layers on performance; evaluate the effectiveness of the attention-based fusion mechanism compared to simpler fusion techniques ; and evaluate the impact of multi-task learning. Table III presents the results.

The self-attention mechanism plays a pivotal role in capturing inter-modal dependencies, allowing the model to effectively generalize across diverse datasets. Its ability to focus on relevant features across multiple modalities ensures robust performance in complex scenarios. Furthermore, experiments reveal that reducing the number of transformer layers results in significant performance degradation, underscoring the critical importance of depth in accurately modeling the intricate relationships between modalities. The attention-based fusion method plays a crucial role in integrating modality-specific features by dynamically weighing the contributions of each modality based on their relevance to the emotion recognition task. Unlike simpler fusion techniques, such as direct concatenation or averaging, which treat all modalities equally and may overlook intricate interdependencies, the attention mechanism selectively emphasizes the most informative features from each modality. This results in a more effective and adaptive fusion process, leading to significantly improved performance in group emotion recognition. The multi-task learning method outperforms single-task learning by up to 4% on IEMOCAP and MELD, confirming that leveraging interpersonal relationships with individual emotion predictions enhances GER.

TABLE III
IMPORTANCE OF ARCHITECTURAL COMPONENTS. WITHOUT SELF ATTENTION (W/OSA), REDUCED TRANSFORMER 3-LAYERS (RT3L), FULL TRANSFORMER 6-LAYERS (FT6L); AVERAGING (AVG), CONCATENATION (CONC), ATTENTION (ATT); STL (SINGLETASK LEARNING), MTL (MULTITASK LEARNING)

	IEMOCAP		MELD		VGAF	
Criteria	**Acc**	**F1**	**Acc**	**F1**	**Acc**	**F1**
W/o SA	56.69	56.01	55.36	54.08	49.08	47.73
RT3L	64.43	62.94	62.91	61.67	64.58	62.89
FT6L	**70.03**	**68.41**	**68.92**	**67.60**	**72.56**	**71.47**
Avg	54.62	52.67	53.54	51.82	54.42	52.17
Conc	59.52	58.01	56.31	54.97	56.59	55.53
Att	**70.03**	**68.41**	**68.92**	**67.60**	**72.56**	**71.47**
STL	70.03	68.41	68.92	67.60	72.56	71.47
MTL	**74.91**	**72.38**	**72.95**	**71.57**	**72.56**	**71.47**

3) Comparison with State-of-the-Arts: To evaluate the effectiveness of the proposed framework, we conduct a comprehensive ablation study in table IV focusing on the impact of key components through the F1-score: the multimodal transformer architecture, hierarchical feature aggregation, multitask learning (MTL), and a comparison with a simple feature concatenation baseline. Additionally, we include a direct comparison with [1], [9], [11], a recent state-of-the-arts model for Emotion Recognition (GER). The comparison with state-of-the-arts shows that the proposed framework outperforms SCFA and VGAFNet across all datasets, achieving state-of-the-art results. Its advanced feature extraction, effective fusion mechanism, and multimodal transformer architecture enable robust modeling of inter-modal dependencies, ensuring high accuracy and reliability in group emotion recognition.

TABLE IV
COMPONENT-WISE EVALUATION

Model Variant	MELD	IEMOCAP	VGAF
W/o MTL	70.6	73.2	70.1
W/o Hierarchical aggragation	68.9	70.7	68.0
W/o Transformer (simple concat)	65.8	68.4	66.1
[9]	63.69	66.42	-
[11]	-	-	71.16
[1]	70.1	76.2	65.7
Full Framework (Ours)	**72.4**	**75.6**	**72.1**

- Multimodal Transformer + Aggregation: Removing the hierarchical aggregation leads to a substantial drop, especially in MELD and IEMOCAP, indicating its effectiveness in modeling inter-personal and inter-modal dependencies.
- MTL: The absence of multitask learning causes a performance drop (approx. -1.8 F1 on MELD), validating that joint learning of group and individual emotion benefits both tasks.
- Simple Concatenation: Using simple feature concatenation significantly underperforms across all datasets, underscoring the need for structured attention-based fusion.
- [1] vs. Our Framework: While [1] slightly outperforms us on IEMOCAP (due to its tailored hierarchical encoding for dyadic interactions), our framework delivers superior results on MELD and VGAF, thanks to its scalable multimodal design and broader task generalization.

This study confirms that each component plays a critical role in the success of the framework. Regarding transformer depth, reducing the number of layers caused significant performance drops, highlighting its importance in modeling complex dependencies. Similarly, removing inter- or intra-modal self-attention mechanisms led to degraded accuracy, underscoring their necessity for capturing both inter- and intra-modal relationships. For hierarchical aggregation, replacing attention-based pooling with simple averaging reduced performance, demonstrating the value of modeling interpersonal dynamics. Our hierarchical fusion strategy was also compared to a baseline that merely concatenates features from all individuals; results showed that hierarchical fusion substantially outperforms this naive approach, emphasizing the effectiveness of our design. Furthermore, eliminating the group-level prediction task and training solely for individual emotion recognition caused a performance decline (1.4% drop in F1-score), indicating that multitask learning enables shared representations that benefit both tasks. While the framework introduces moderate computational overhead—primarily from its multi-branch encoders and hierarchical fusion—the transformer blocks account for most of the cost, with complexity $O(n^2d)$ where n is the multimodal sequence length and d the hidden dimension. Nonetheless, modality-specific encoders and hierarchical aggregation mitigate unnecessary cross-modal interactions, and the multitask head adds minimal overhead by reusing shared parameters.

V. Conclusion

This paper introduced a novel multimodal framework for group emotion recognition that leverages hierarchical feature aggregation and a transformer-based architecture. By modeling intra- and inter-modal dependencies through self-attention mechanisms and employing multitask learning for joint individual and group emotion prediction, the framework achieves superior performance across diverse datasets. The inclusion of a learned separator token and attention masking strategies enables scalable handling of variable group sizes. Extensive experiments and ablation studies validate the effectiveness of each component, demonstrating the advantages of our approach over existing baselines such as SCFA and VGAFNet. The framework's flexibility, scalability, and robust performance position it as a promising solution for real-world applications in group affective computing. Future work will focus on enhancing temporal modeling and extending to spontaneous multimodal group interactions in dynamic environments.

References

[1] Morais, Edmilson, Ron Hoory, Weizhong Zhu, Itai Gat, Matheus Damasceno, and Hagai Aronowitz. "Speech emotion recognition using self-supervised features." In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6922-6926. IEEE, 2022

[2] Li, Sunan, Hailun Lian, Cheng Lu, Yan Zhao, Chuangao Tang, Yuan Zong, and Wenming Zheng. "Audio-Visual Group-based Emotion Recognition using Local and Global Feature Aggregation based Multi-Task Learning." In Proceedings of the 25th International Conference on Multimodal Interaction, pp. 741-745. 2023.

[3] Tan, Lianzhi, Kaipeng Zhang, Kai Wang, Xiaoxing Zeng, Xiaojiang Peng, and Yu Qiao. "Group emotion recognition with individual facial emotion CNNs and global image based CNNs." In Proceedings of the 19th ACM international conference on multimodal interaction, pp. 549-552. 2017.

[4] Kaviya, P., and T. Arumugaprakash. "Group facial emotion analysis system using convolutional neural network." In 2020 4th International conference on trends in electronics and informatics (ICOEI)(48184), pp. 643-647. IEEE, 2020.

[5] Veltmeijer, Emmeke A., Charlotte Gerritsen, and Koen V. Hindriks. "Automatic emotion recognition for groups: a review." IEEE Transactions on Affective Computing 14, no. 1 (2021): 89-107.

[6] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." Advances in neural information processing systems 30 (2017).

[7] Fujii, Katsuya, Daisuke Sugimura, and Takayuki Hamamoto. "Hierarchical group-level emotion recognition." IEEE Transactions on Multimedia 23 (2020): 3892-3906.

[8] Busso, Carlos, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. "IEMOCAP: Interactive emotional dyadic motion capture database." Language resources and evaluation 42 (2008): 335-359.

[9] Zhao, Huan, Bo Li, and Zixing Zhang. "Speaker-aware cross-modal fusion architecture for conversational emotion recognition." In Proc. Interspeech, pp. 2718-2722. 2023.

[10] Poria, Soujanya, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. "Meld: A multimodal multi-party dataset for emotion recognition in conversations." arXiv preprint arXiv:1810.02508 (2018).

[11] Sharma, Garima, Abhinav Dhall, and Jianfei Cai. "Audio-visual automatic group affect analysis." IEEE Transactions on Affective Computing 14, no. 2 (2021): 1056-1069.

[12] Nagarajan, Bhalaji, and V. Ramana Murthy Oruganti. "Group emotion recognition in adverse face detection." In 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), pp. 1-5. IEEE, 2019.

[13] Kim, Taewoon, and Piek Vossen. "Emoberta: Speaker-aware emotion recognition in conversation with roberta." arXiv preprint arXiv:2108.12009 (2021).

[14] Huang, Xiaohua, Jinke Xu, Wenming Zheng, Qirong Mao, and Abhinav Dhall. "A Survey of Deep Learning for Group-level Emotion Recognition." arXiv preprint arXiv:2408.15276 (2024).

[15] Sharma, Garima, Shreya Ghosh, and Abhinav Dhall. "Automatic group level affect and cohesion prediction in videos." In 2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), pp. 161-167. IEEE, 2019.

[16] Fujii, Katsuya, Daisuke Sugimura, and Takayuki Hamamoto. "Hierarchical group-level emotion recognition in the wild." In 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), pp. 1-5. IEEE, 2019.

[17] Guo, Xin, Luisa F. Polanía, and Kenneth E. Barner. "Group-level emotion recognition using deep models on image scene, faces, and skeletons." In Proceedings of the 19th ACM International Conference on Multimodal Interaction, pp. 603-608. 2017.

[18] Guo, Xin, Bin Zhu, Luisa F. Polanía, Charles Boncelet, and Kenneth E. Barner. "Group-level emotion recognition using hybrid deep models based on faces, scenes, skeletons and visual attentions." In Proceedings of the 20th ACM international conference on multimodal interaction, pp. 635-639. 2018.

[19] Khan, Ahmed Shehab, Zhiyuan Li, Jie Cai, Zibo Meng, James O'Reilly, and Yan Tong. "Group-level emotion recognition using deep models with a four-stream hybrid network." In Proceedings of the 20th ACM international conference on multimodal interaction, pp. 623-629. 2018.

[20] LI, Sunan, LIAN, Hailun, LU, Cheng, et al. "Audio-Visual Group-based Emotion Recognition using Local and Global Feature Aggregation based Multi-Task Learning". In : Proceedings of the 25th International Conference on Multimodal Interaction. p. 741-745. 2023.