Emot-CM-BERT: Adaptive Attention and Class-Aware Cross-Modal Learning for Emotion Recognition from Audio and Text

Shintami Chusnul Hidayati, James Rafferty Lee, and Kevin Davi Samuel
Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
E-mail: shintami@its.ac.id

***Abstract*—Emotion recognition from multimodal data has become increasingly vital for affective computing applications such as virtual assistants, social robots, and mental health monitoring. Among various modalities, audio and text offer complementary emotional signals, but effectively integrating them remains a challenge due to asynchronous signals, modality imbalance, and static fusion strategies. This paper presents Emot-CM-BERT, a cross-modal framework designed to recognize emotion from audio and text inputs. The model integrates acoustic features—such as intonation, pitch, and rhythm—with semantic features that capture the contextual meaning of language, using a masked multimodal attention mechanism that dynamically fuses information across modalities. To further optimize learning, curriculum learning is employed to progressively introduce training complexity, while class-aware sampling addresses imbalances in emotion label distributions. We evaluate Emot-CM-BERT on three benchmark datasets—CREMA-D, IEMOCAP, and CMU-MOSEI—encompassing both acted and naturalistic emotional expressions. Experimental results demonstrate that the proposed model achieves competitive or superior performance compared to baselines, particularly in real-world, conversational settings. The analysis further confirms its effectiveness in aligning heterogeneous features and generalizing across diverse emotions.**

I. Introduction

Emotion recognition is a key task in affective computing that enables machines to interpret users' emotional states in a way that mimics human interaction. It has become increasingly important in human-computer interaction, supporting applications such as virtual assistants, social robots, e-learning systems, and mental health monitoring tools [1]. As multimodal data becomes more available, integrating audio and text has become a prominent strategy to capture semantic and paralinguistic signals for understanding emotions. Text provides semantic context and sentiment, while audio adds prosodic cues like pitch and intonation. Moreover, the growing use of multimedia on social platforms highlights the importance of contextual emotional understanding, where not only content but also user profile and metadata influence interaction and popularity [2].

Recent advances in deep learning have improved the performance of emotion recognition systems. Transformer-based models such as BERT [3] have revolutionized natural language processing by capturing deep contextual dependencies. Concurrently, audio analysis has benefited from convolutional and recurrent neural networks, which model temporal and spectral characteristics [4], [5]. These advances have inspired the development of multimodal fusion architectures that integrate features across modalities for better emotional nuances.

Various multimodal models have explored early, late, and hybrid fusion strategies to combine audio and textual features. Yoon *et al.* [6] showed that feature-level fusion improves emotional inference, whereas Sahu [7] adopted late fusion to reduce noise caused by cross-modal inconsistencies. More sophisticated models, such as hierarchical attention networks [8] and ensemble-based architectures [9], aim to better capture the interdependencies between modalities.

With the rise of attention mechanisms, transformer-based architectures have become the standard for multimodal learning. CM-BERT [10], for example, extends BERT [3] to support cross-modal tasks by enabling mutual attention between text and audio modalities. Similarly, the multimodal transformer [11] and its derivatives introduced aligned and unaligned attention mechanisms to address the asynchronous characteristics inherent in multimodal signals. More recently, transformer autoencoders have been applied to learn cross-modal correlations specially for emotion recognition tasks [12].

Several recent studies further highlight the promise of these architectures. A study by Kakuba *et al.* [13] analyzes deep multimodal architectures and shows their superiority in leveraging complementary features across audio and text. Additionally, Yang *et al.* [14] introduced curriculum learning to improve the training stability of sequence models in conversation, a technique that has shown potential in multimodal emotion tasks. Contrastive learning has also been used to extract modality-invariant representations, allowing more generalizable emotion recognition across datasets and domains [15].

Nonetheless, several challenges persist in effectively modeling emotions from audio and text. First, emotional cues are often unevenly distributed across modalities, yet many existing approaches rely on static or manually structured fusion strategies that fail to adaptively prioritize the most informative signals in a given context [13]. Second, most models are designed for unimodal settings or assume fixed modality contributions, lacking the ability to dynamically adjust modality importance based on contextual relevance [14], [15]. This limits their generalization across diverse emotional expressions, especially in real-world, multimodal scenarios. Third, emotion datasets are typically imbalanced. Without class-aware mechanisms,

979-8-3315-7206-8/25/$31.00

models tend to be biased toward majority classes [15].

Prior works in multimedia understanding highlight the value of contextual and perceptual cues. Image popularity prediction leverages category-specific and user interaction features [16], while body shape estimation from anthropometric data supports personalized modeling [17]. In aesthetic analysis, combining structural and color features across multiple color spaces has improved subjective evaluations, such as in food imagery [18], [19]. These findings suggest that emotion recognition can similarly benefit from context-aware, fine-grained feature integration, and dynamic representation learning.

To address these limitations, we propose *Emot-CM-BERT*, a modified CM-BERT-based framework specifically designed for emotion recognition from audio and text. Our approach introduces three key innovations: (1) an expanded input layer that integrates Covarep-extracted acoustic features with BERT-based text embeddings, (2) a masked multimodal attention mechanism that facilitates dynamic cross-modal fusion by allowing one modality to guide attentional focus in the other, and (3) a classification head optimized for categorical emotion identification. To further improve model performance and robustness, we incorporate curriculum learning and class-weighted sampling, which stabilize training and mitigate the impact of class imbalance.

We evaluated Emot-CM-BERT on three widely used benchmark datasets, covering both acted and naturalistic emotional expressions. Experimental results demonstrate that the proposed model achieves competitive or superior performance compared to baselines, particularly in realistic conversational contexts. Detailed analysis further confirms its effectiveness in aligning heterogeneous characteristics and adapting to diverse emotional conditions.

Our main contributions are summarized as follows.

1) We propose a novel cross-modal architecture for emotion recognition, featuring a masked multimodal attention mechanism for adaptive audio-text fusion.
2) We introduce a curriculum-based training strategy combined with class-weighted sampling to improve training stability and address class imbalance.
3) We conduct comprehensive evaluations on three benchmark datasets, demonstrating the effectiveness and generalizability of the proposed model in both controlled and real-world emotional settings.

The rest of this paper is organized as follows. Section II describes the methodology, including the architecture of the proposed Emot-CM-BERT and its key adaptations. Section III details the experimental setup and provides an analysis of the model performance compared to baseline methods. Finally, Section IV summarizes the contributions and discusses future research directions.

II. Methodology

This section presents the proposed Emot-CM-BERT framework for multimodal emotion recognition. It begins with the extraction of modality-specific features from audio and text, followed by a description of architectural modifications to the original CM-BERT [10].

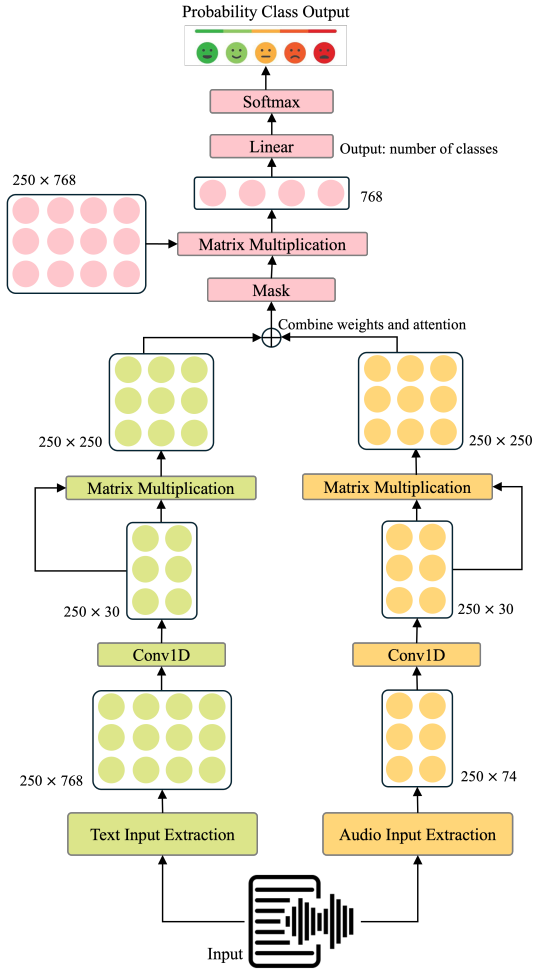

Fig. 1. Architecture of the proposed Emot-CM-BERT model.

A. Feature Extraction

This study leverages two complementary modalities—audio and text—to capture emotional expressions, each processed through modality-specific feature extraction techniques. For audio, we utilize Covarep [20], an open-source toolkit that extracts prosodic and spectral features such as pitch, voicing probability, and harmonic descriptors. These features capture temporal and tonal properties for inferring emotional states.

For the textual modality, we employ BERT [3] to generate deep contextual embeddings. By incorporating bidirectional context, BERT models the semantic nuances of each word within a sentence. Each modality is processed independently to extract discriminative features before being integrated into a unified multimodal representation.

B. Model Architecture

CM-BERT [10] is a cross-modal extension of BERT that facilitates mutual attention between modalities by employing a

TABLE I
PERFORMANCE COMPARISON BETWEEN THE PROPOSED METHOD AND BASELINE MODELS.

Dataset	Method	Accuracy	Precision	Recall	F1-score
CREMA-D	Sahu [7]	0.4730	0.4740	0.4730	0.4570
	Yoon *et al.* [6]	**0.4859**	**0.4942**	**0.4859**	**0.4692**
	Emot-CM-BERT (ours)	0.4671	0.4760	0.4671	0.4669
IEMOCAP	Sahu [7]	0.5710	0.2980	0.2340	0.2380
	Yoon *et al.* [6]	0.5580	0.5578	0.5580	0.5572
	Emot-CM-BERT (ours)	**0.5932**	**0.6024**	**0.5932**	**0.5912**
CMU-MOSEI	Sahu [7]	0.2800	0.2240	0.1960	0.2000
	Yoon *et al.* [6]	0.5522	**0.5596**	0.5522	0.5517
	Emot-CM-BERT (ours)	**0.6412**	0.5425	**0.6412**	**0.5715**

shared transformer backbone. Modality-specific inputs are encoded jointly through self-attention layers, enabling contextual interactions across modalities. Although effective for general cross-modal tasks, the original CM-BERT architecture was primarily designed for regression objectives and lacks direct support for high-dimensional acoustic features commonly used in emotion recognition.

As shown in Fig. 1, we introduce several modifications to the CM-BERT framework to better suit emotion recognition tasks. First, the input layer is adapted to incorporate Covarep-based acoustic features alongside BERT-based textual embeddings. This allows the model to handle the higher-dimensional audio data not originally handled by CM-BERT.

Next, we reconfigure the output layer to perform emotion classification. The original CM-BERT output, designed for regression tasks, is replaced with a classification head that outputs probability distributions over discrete emotion labels– better aligning with standard emotion recognition datasets.

We integrate a masked multimodal attention mechanism to fuse audio and textual modalities. This mechanism facilitates cross-modal alignment by allowing the audio modality to guide attention within the textual domain. Such an approach is particularly beneficial when emotional cues are more prominent in one modality.

To further optimize model performance, we employ several training strategies. Curriculum learning is employed to present training samples in a structured manner—starting from simpler examples and gradually progressing to more complex ones—to support stable learning. Additionally, to address class imbalance commonly observed in emotion datasets, we apply a weighted sampling strategy that increases the frequency of underrepresented emotional classes during training.

III. EXPERIMENTAL RESULTS

A. Dataset

We evaluate the proposed model using three widely adopted benchmark datasets: CMU-MOSEI [21], IEMOCAP [22], and CREMA-D [23]. These datasets captures emotions across both natural and controlled settings.

CMU-MOSEI: The CMU-MOSEI dataset is a large-scale dataset consisting of over 23,000 video segments extracted from online monologues. Each segment is annotated with six emotion categories. Its diverse, realistic expressions make it ideal for evaluating models in unconstrained settings.

IEMOCAP: The IEMOCAP dataset comprises approximately 12 hours of scripted and improvised dyadic conversations between actors. Its focus on interactive and emotional exchanges makes it particularly relevant for evaluating performance in conversational settings.

CREMA-D: CREMA-D comprises over 7,000 audio-visual clips of professional actors expressing six predefined emotions under controlled conditions. This dataset serves as a useful benchmark for assessing the model's ability to detect clearly articulated emotional signals.

B. Comparative Methods

To assess the effectiveness of the proposed Emot-CM-BERT model, we compare its performance against two representative multimodal emotion recognition methods.

Yoon et al. *[6]:* This approach utilizes deep learning-based feature-level fusion to integrate audio and textual inputs. This method serves as a strong baseline for evaluating the impact of our cross-modal attention mechanism.

Sahu [7]: This method addresses ambiguity in multimodal data by extracting modality-specific characteristics followed by late fusion. Its focus on inter-modal inconsistencies offers a relevant benchmark for evaluating our model's ability to align and harmonize heterogeneous features from audio and text.

C. Results and Discussion

The experimental results in Table I demonstrate how dataset-specific characteristics influence the performance of the proposed model relative to the baseline methods. On the CREMA-D dataset, the proposed method achieved an accuracy of 0.4671, below the best-performing baseline by Yoon *et al.* at 0.4859. Similarly, precision and F1-score differences were marginal, suggesting a misalignment between the model's adaptive mechanisms and the structured, acted nature of CREMA-D. This gap may stem from the dataset's scripted expressions, which differ from the naturalistic contexts for which the model was optimized, thus limiting its generalization.

In contrast, the proposed method achieved the highest performance on the IEMOCAP dataset, achieving an accuracy of 0.5932 and outperforming both Sahu (0.5710) and Yoon *et al.* (0.5580). Improvements were also observed in precision, recall, and F1-score. IEMOCAP's dialog-based, naturalistic interactions align well with the model's design—particularly

TABLE II
IMPACT OF DIFFERENT MODALITY ON THE PERFORMANCE OF THE PROPOSED METHOD.

Dataset	Modality	Accuracy	Precision	Recall	F1-score
CREMA-D	Audio	**0.4671**	**0.4760**	**0.4671**	**0.4669**
	Text	0.1718	0.0602	0.1718	0.0834
	Text + Audio	0.1603	0.1300	0.1603	0.1315
IEMOCAP	Audio	0.2873	0.2627	0.2873	0.2677
	Text	0.5348	0.5588	0.5348	0.5243
	Text + Audio	**0.5932**	**0.6024**	**0.5932**	**0.5912**
CMU-MOSEI	Audio	0.5693	0.4238	0.5693	0.4370
	Text	0.5787	0.4796	0.5787	0.5197
	Text + Audio	**0.6412**	**0.5425**	**0.6412**	**0.5715**

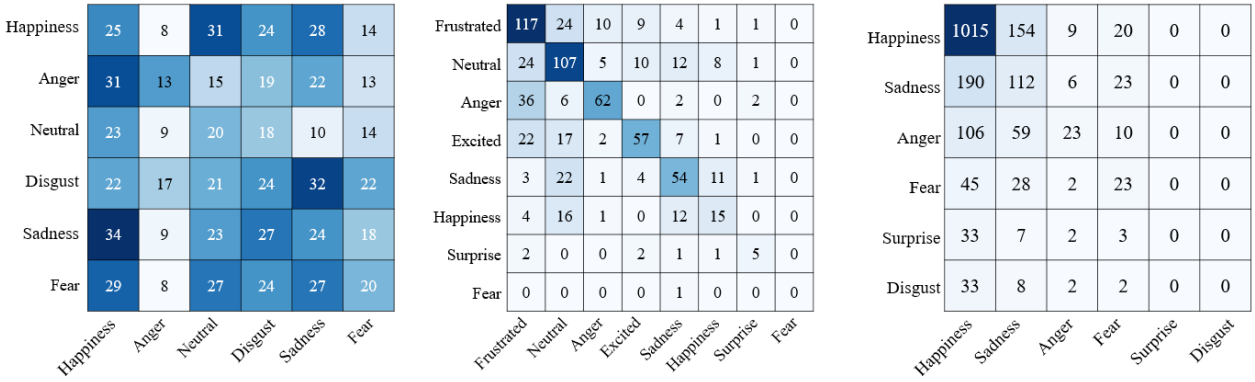

Fig. 2. Confusion matrices of the proposed method across three benchmark datasets–CREMA-D, IEMOCAP, and CMU-MOSEI (left to right). The vertical axis denotes actual classes; the horizontal axis denotes predicted classes.

its cross-modal attention mechanisms that effectively leverage the interplay between audio and text features.

In the CMU-MOSEI dataset, the proposed method also performed strongly, attaining the highest accuracy (0.6412) and F1-score (0.5715) among all models. While Yoon *et al.* reported slightly higher precision (0.5596 vs. 0.5425), our model exhibited better recall, indicating improved sensitivity to emotional cues. The strong performance is attributed to CMU-MOSEI's diverse and realistic multimodal content, which aligns well with our model's feature fusion strategy. The modest decline in precision suggests a tendency to overgeneralize in some cases, which could be mitigated by future refinement of the attention mechanism to enhance specificity.

Table II summarizes the performance impact of individual and combined modalities. Across most datasets, the multimodal setting consistently produced the best results, confirming the benefit of integrating complementary features. An exception is CREMA-D, where combining modalities reduced performance, suggesting that textual features introduced noise or inconsistencies.

Specifically, on CREMA-D, audio alone achieved the highest accuracy (0.4671) and F1-score (0.4669), while text-only inputs performed poorly (accuracy: 0.1718, F1-score: 0.0834). Combining both modalities further degraded performance (accuracy: 0.1603, F1-score: 0.1315), likely due to limited semantic variability in the scripted text may have introduced noise or confusion during multimodal fusion. These findings highlight the need for more adaptive modality-specific weighting mechanisms in cases where one modality is dominant.

For IEMOCAP, text outperformed audio (accuracy: 0.5348 vs. 0.2873; F1-score: 0.5243 vs. 0.2677), yet their combination produced the best performance (accuracy: 0.5932; F1-score: 0.5912). Although audio alone contributed less, its integration with text enhanced the model's overall capability to interpret multimodal emotional cues, validating the effectiveness of our attention-based fusion approach.

On CMU-MOSEI, text and audio performed comparably when used individually, with text yielding slightly higher accuracy (0.5787 vs. 0.5693) and F1-score (0.5197 vs. 0.4370). Their combination improved both metrics (accuracy: 0.6412; F1-score: 0.5715), underscoring the model's ability to exploit complementary features across modalities.

Fig. 2 presents the confusion matrices for the three datasets. For CREMA-D, the matrix shows a widespread misclassification with a weak diagonal trend, particularly among *Fear*, *Sadness*, and *Disgust*. This indicates difficulty in distinguishing between acted emotional expressions. In contrast, the IEMOCAP matrix shows stronger diagonal dominance, especially for *Frustrated*, *Neutral*, and *Sadness*, indicating improved performance in natural conversational contexts. Nevertheless, confusion remains between semantically similar emotions such as *Excited* and *Happy*. The CMU-MOSEI matrix demonstrates the highest degree of class separation, with particularly strong precision for *Happiness*, confirming the model's in handling diverse, in-the-wild emotional content enriched by nuanced linguistic and acoustic cues.

Fig. 3 presents the bootstrapped distribution of F1 scores

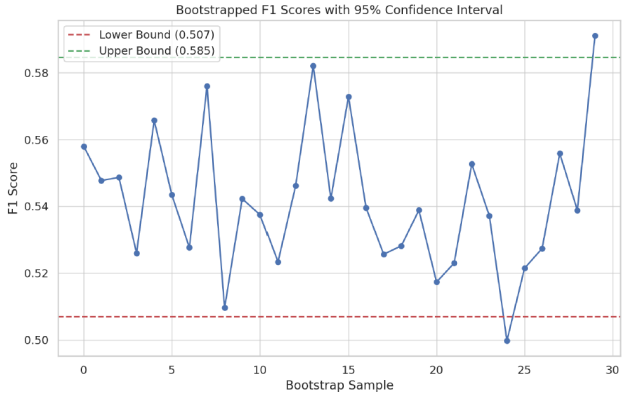

Fig. 3. Bootstrapped F1 score distribution with 95% confidence interval on the IEMOCAP dataset.

on the IEMOCAP dataset, which achieved the best overall performance. The 95% confidence interval, derived from 30 bootstrap iterations, ranges from 0.507 to 0.585 (indicated by red and green dashed lines, respectively).

The lower bound of the confidence interval exceeds 0.500, offering strong statistical evidence that the model performs better than random guessing. Given that a uniform random classifier over eight emotion classes would yield an expected F1-score of approximately 0.125, this result affirms the robustness and generalization capability of the proposed model under diverse sampling conditions and within naturalistic emotional settings.

D. Ablation Analysis

Table III presents an ablation study that examines the contributions of class weighting and curriculum learning to the performance of the proposed Emot-CM-BERT framework. The analysis is conducted on the IEMOCAP and CMU-MOSEI datasets, where the model shows strong baseline performance.

On the IEMOCAP dataset, the full Emot-CM-BERT model achieves the highest overall performance, with an accuracy of 0.5932 and an F1-score of 0.5912. When both CW and CL are removed, performance drops slightly to 0.5789 in accuracy and 0.5778 in F1-score. This decline suggests that CW and CL contribute to stabilizing training and improving generalization. When replaced with data-level techniques—random oversampling (ROS) and random undersampling (RUS)—performance declines further (accuracy: 0.5462, F1-score: 0.5391). These findings indicate that while sampling techniques can provide basic balancing, they are less effective than model-aware strategies in learning from imbalanced emotional data.

A comparable pattern is observed on the CMU-MOSEI dataset. The full model again yields the best accuracy (0.6412) and recall (0.6412), with a F1-score of 0.5715. Removing CW and CL reduces the performance to 0.5920 accuracy and 0.5322 F1-score. Interestingly, the ROS and RUS variant achieve the highest precision (0.5719), slightly exceeding the full model (0.5425). This suggests that while sampling may improve per-class discrimination, it does not support balanced performance across metrics, particularly recall and F1-score, which are critical in emotion classification tasks involving imbalanced labels.

In summary, the results confirm that class weighting and curriculum learning are integral to the effectiveness of Emot-CM-BERT, particularly in scenarios where the emotional class distribution is skewed or when learning from variable-complexity inputs. Although sampling-based methods can offer localized improvements in specific metrics (e.g., precision), they do not provide the same level of consistency and robustness across evaluation dimensions as the complete model configuration.

IV. Conclusion and Future Work

This study proposed Emot-CM-BERT, a modified cross-modal transformer framework for multimodal emotion recognition that integrates audio and textual information. By incorporating masked multimodal attention, curriculum learning, and class-aware sampling, the model effectively captures complementary emotional cues across modalities and achieves competitive performance compared to state-of-the-art approaches.

Experimental results highlight the importance of dataset characteristics in determining modality contributions. Emot-CM-BERT performs particularly well on naturalistic datasets, where audio and text modalities offer rich and complementary emotional signals. However, its performance is comparatively limited on controlled dataset, where modality imbalance and scripted expressions may reduce the effectiveness of cross-modal fusion.

For future work, we plan to explore adaptive modality-specific weighting strategies to better address modality imbalance across varying contexts. Additionally, we aim to extend the framework by incorporating visual features to support more comprehensive multimodal emotion recognition.

Acknowledgement

This work was partially supported by Institut Teknologi Sepuluh Nopember under Grant Nos. 1019/PKS/ITS/2024 and 2309/PKS/ITS/2025.

References

[1] C. Singla, S. Singh, P. Sharma, N. Mittal, and F. Gared, "Emotion recognition for human–computer interaction using high-level descriptors," *Scientific Reports*, 2024.

[2] S. C. Hidayati, M. R. Fiqih Thalib, and A. Munif, "The influence of user profile and post metadata on the popularity of image-based social media: A data perspective," in *Proc. Int. Conf. Artificial Intelligence in Information and Communication*, 2024, pp. 806–811.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, J. Burstein, C. Doran, and T. Solorio, Eds., 2019, pp. 4171–4186.

[4] A. Mehrish, N. Majumder, R. Bharadwaj, R. Mihalcea, and S. Poria, "A review of deep learning techniques for speech processing," *Information Fusion*, vol. 99, no. C, 2023.

TABLE III
ABLATION STUDY ON THE IMPACT OF CLASS WEIGHTING (CW) AND CURRICULUM LEARNING (CL) IN THE PROPOSED EMOT-CM-BERT FRAMEWORK. ROS = RANDOM OVERSAMPLING, RUS = RANDOM UNDERSAMPLING.

Dataset	Method	Accuracy	Precision	Recall	F1-score
IEMOCAP	Emot-CM-BERT w/o CW & CL	0.5789	0.5869	0.5789	0.5778
	Emot-CM-BERT w/o CW & CL, with ROS & RUS	0.5462	0.5552	0.5462	0.5391
	Emot-CM-BERT (Full Model)	**0.5932**	**0.6024**	**0.5932**	**0.5912**
CMU-MOSEI	Emot-CM-BERT w/o CW & CL	0.5920	0.5127	0.5920	0.5322
	Emot-CM-BERT w/o CW & CL, with ROS & RUS	0.5523	**0.5719**	0.5523	0.5571
	Emot-CM-BERT (Full Model)	**0.6412**	0.5425	**0.6412**	**0.5715**

[5] S. C. Hidayati, A. Satria Adidarma, and K. R. Sungkono, "Exploring the impact of spatio-temporal patterns in audio spectrograms on emotion recognition," in *Proc. Int. Conf. Advanced Mechatronics, Intelligent Manufacture and Industrial Automation*, 2023, pp. 200–205.

[6] S. Yoon, S. Byun, and K. Jung, "Multimodal speech emotion recognition using audio and text," in *Proc. IEEE Spoken Language Technology Workshop*, 2018, pp. 112–118.

[7] G. Sahu, "Multimodal speech emotion recognition and ambiguity resolution," *arXiv preprint arXiv:1904.06022*, 2019.

[8] Y. Gu, K. Yang, S. Fu, S. Chen, X. Li, and I. Marsic, "Multimodal affective analysis using hierarchical attention strategy with word-level alignment," in *Proc. Annual Meeting of the Association for Computational Linguistics*, 2018, pp. 2225–2235.

[9] S. C. Hidayati, M. Subhan, and Y. Anistyasari, "A novel stacking ensemble learning approach for emotion detection in audio-to-text transcriptions," in *Proc. Int. Seminar on Intelligent Technology and Its Applications*, 2024, pp. 512–517.

[10] K. Yang, H. Xu, and K. Gao, "Cm-bert: Cross-modal bert for text-audio sentiment analysis," in *Proc. ACM Int. Conf. Multimedia*, 2020, pp. 521–528.

[11] Y.-H. H. Tsai, S. Bai, P. P. L. Yamada, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proc. Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 6558–6569.

[12] C. Cheng, W. Liu, Z. Fan, L. Feng, and Z. Jia, "A novel transformer autoencoder for multi-modal emotion recognition with incomplete data," *Neural Networks*, vol. 172, no. C, 2024.

[13] S. Kakuba, A. Poulose, and D. S. Han, "Deep learning-based speech emotion recognition using multi-level fusion of concurrent features," *IEEE Access*, vol. 10, pp. 125 538–125 551, 2022.

[14] L. Yang, Y. Shen, Y. Mao, and L. Cai, "Hybrid curriculum learning for emotion recognition in conversation," in *Proc. AAAI Conf. Artificial Intelligence*, vol. 36, 2022, pp. 11 595–11 603.

[15] K. Yang, T. Zhang, H. Alhuzali, and S. Ananiadou, "Cluster-level contrastive learning for emotion recognition in conversations," *IEEE Trans. Affective Computing*, vol. 14, no. 4, pp. 3269–3280, 2023.

[16] E. Massip, S. C. Hidayati, W.-H. Cheng, and K.-L. Hua, "Exploiting category-specific information for image popularity prediction in social media," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, 2018, pp. 45–46.

[17] S. C. Hidayati and Y. Anistyasari, "Body shape calculator: Understanding the type of body shapes from anthropometric measurements," in *Proc. Int. Conf. Multimedia Retrieval*, 2021, pp. 461–465.

[18] S. C. Hidayati, M. Valda Rizky Nur Firdaus, R. W. Nur Dianto, and Sarwosri, "Unleashing attributes-content adaptation with multi-color spaces for food photo aesthetic assessment," in *Proc. Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, 2024, pp. 1–6.

[19] S. C. Hidayati, M. A. Ardiansyah, Sarwosri, Y. Anistyasari, and W.-H. Cheng, "A hybrid approach to food image aesthetic assessment via structural and color-based feature integration," in *Proc. Int. Conf. Consumer Technology*, 2025, pp. 1–4.

[20] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "Covarep — a collaborative voice analysis repository for speech technologies," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2014, pp. 960–964.

[21] A. Bagher Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *Proc. Annual Meeting of the Association for Computational Linguistics*, 2018, pp. 2236–2246.

[22] C. Busso, M. Bulut, C.-C. Lee, *et al.*, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, 2008.

[23] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE Trans. Affective Computing*, vol. 5, no. 4, pp. 377–390, 2014.