

# MERSA: Multimodal Emotion Recognition with Self-Align Embedding

Quan Bao Le 

*Dept. of IT Specialization*

*FPT University*

Ho Chi Minh City, Vietnam

quanlbse160758@fpt.edu.vn


Kiet Tuan Trinh 

*Dept. of IT Specialization*

*FPT University*

Ho Chi Minh City, Vietnam

kietttse160148@fpt.edu.vn

Nguyen Dinh Hung Son 

*Dept. of IT Specialization*

*FPT University*

Ho Chi Minh City, Vietnam

sonndhse160761@fpt.edu.vn

Phuong-Nam Tran 

*Dept. of Computing Fundamental*

*FPT University*

Ho Chi Minh City, Vietnam

namtpse150004@fpt.edu.vn

Cuong Tuan Nguyen 

*Faculty of Engineering*

*Vietnamese-German University*

Ho Chi Minh City, Vietnam

cuong.nt2@vgu.edu.vn

Duc Ngoc Minh Dang\* 

*Dept. of Computing Fundamental*

*FPT University*

Ho Chi Minh City, Vietnam

ducnm2@fe.edu.vn

**Abstract**—Emotions are an integral part of human communication and interaction, significantly shaping our social connections, decision-making, and overall well-being. Understanding and analyzing emotions have become essential in various fields, including psychology, human-computer interaction, marketing, and healthcare. The previous approach has indeed made significant strides in improving the accuracy of predicting emotions within speech. However, the current model's performance still falls short when it comes to real-life applications. This limitation arises due to several factors such as lack of context, ambiguity in speech and meaning, and other contributing elements. To reduce the ambiguity of emotions within speech, this paper seeks to leverage multiple data modalities, specifically textual and acoustic information. To analyze these modalities, we propose a novel approach called MERSA which utilizes the self-align method to extract context features from both textual and acoustic information. By leveraging this technique, the MERSA model can effectively create fusion feature vectors of the multiple inputs, facilitating a more accurate and holistic analysis of emotions within speech. Moreover, the MERSA model has incorporated a cross-attention module into its network architecture, which enables the MERSA model to capture and leverage the interdependencies between the textual and acoustic modalities.

**Index Terms**—speech emotion recognition, multimodal emotion recognition, self-align embedding, cross-modality attention

## I. INTRODUCTION

Traditional studies of emotion analysis primarily relied on verbal expressions, facial cues, and physiological responses as indicators of emotional states. However, humans convey emotions through multiple channels simultaneously, including speech acoustic, the content of speech, facial expressions, and body language. This multifaceted nature of emotional expression has led to the emergence of Speech Emotion Recognition (SER) and Multimodal Emotion Recognition as potential fields of research.

In the early days of sentiment analysis, rule-based approaches were predominant. These approaches relied on man-

ually crafted rules and heuristics to identify sentiment in text. Research during this phase focused on developing lexicons and dictionaries of sentiment-bearing words, assigning polarity (positive, negative, or neutral) to these words, and using rules to determine sentiment based on word patterns. SER then transitioned from rule-based approaches to machine learning-based methods. This shift allowed for the development of more sophisticated and accurate emotion recognition systems that could learn patterns and features directly from data rather than relying on manually crafted rules. Research during this phase focused on the application of machine learning algorithms for sentiment classification. With the emergence of deep learning, the SER witnessed a significant advancement. Deep neural networks, particularly Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs), began to dominate emotion recognition tasks. Researchers explored the application of RNNs for sequential data, while CNNs were employed for sentiment analysis using convolutional layers to capture local patterns in text or images. Furthermore, researchers began exploring the multimodal method, which involved the fusion of text, audio, and visual data to obtain a more holistic understanding of sentiment. Crossmodal Fusion [1] extends the idea of multimodal fusion by transferring knowledge or representations learned from one modality to another. It leverages information from one modality to improve performance in another modality. Crossmodal fusion can be particularly useful when one modality has more labeled data or is easier to work with than others. It can help improve the performance of emotion recognition or sentiment analysis in modalities with limited data. The main challenge in crossmodal fusion is ensuring that the transferred knowledge is relevant and beneficial for the target modality. It requires careful consideration of feature representations and adaptation techniques.

While previous studies have shown significant advancements, many often overlook inadequately incorporated mul-

\* Corresponding author: Duc Ngoc Minh Dang (ducnm2@fe.edu.vn)

multiple modalities and capture their relation. In 2023, a method called SERVER [2] proposed a strategy to improve the SER model by utilizing both audio and text features. However, SERVER simply concatenates these different types of information together without considering how they are related, which is an important aspect of multimodal analysis. On the other hand, 3M-SER [3] attempted to address the interrelationships among different information by incorporating an attention mechanism and achieved good results. However, 3M-SER still has limitations in its feature extraction strategy. Both SERVER and 3M-SER may introduce additional errors because of the inherent errors in each type of feature and the differences between them.

To address this problem, our research introduces the MERSA model which incorporates self-align embeddings [4] into the feature extractor to fine-tune the text features. Additionally, we employ cross-attention [5] to effectively combine the audio and text features for improved performance. Our primary contributions cover the following three main aspects:

- 1) We adapt the cross-attention for multimodal emotion recognition tasks by effectively aligning and fusing information from different modalities.
- 2) We integrate the self-align embeddings methodology into the BERT-base model. This approach involves transferring knowledge from the other field to the SER field, enabling better representation of learning for SER tasks.
- 3) We conduct a comprehensive assessment and comparison of various models on two well-known datasets: IEMOCAP and MELD. By evaluating the performance of the MERSA model and other models, we gain insights into the effectiveness of our proposed approach over existing methods in emotion recognition tasks.

The rest of this paper is structured as follows. The related work is presented in Section II. The datasets and methodology of our model are presented in Sections III and IV, respectively. In Section V, the preliminary results are displayed and analyzed. Finally, the conclusion and potential future work are concluded and listed in Section VI.

## II. RELATED WORKS

### A. Audio features

Google in 2017 developed a novel approach employing CNN to transform audio signals into a latent space representation called VGGish [6]. VGGish utilizes the log Mel-Spectrogram derived from audio input to extract audio features with each second of an audio file will be transformed into an image of a log Mel-Spectrogram, and then the VGGish networks are applied to serve as a feature extractor. Another approach for audio embeddings is Wav2Vec [7]. Wav2Vec creates speech representations by performing a context prediction task. It utilizes autoencoding to identify discrete units and also learns continuous speech representations through self-supervised context prediction tasks. This method demonstrates the power of self-supervised learning in extracting valuable features from audio data, without the need for labeled datasets. It is particularly effective even with very short audio files.

### B. Text features

Bidirectional Encoder Representations from Transformers (BERT) [8] is a unique model known for its bidirectional nature, allowing it to consider both the preceding and following context of a word in a sentence. It is built upon the transformer architecture [9], which utilizes a self-attention mechanism to determine the importance of different words in a sentence relative to a given word. BERT is pre-trained on a large dataset of text, allowing it to learn contextual feature vectors for words and sentences. BERT is pre-trained on a large dataset of text, allowing it to learn contextual feature vectors for words and sentences. Thus, BERT and its variation are the most used to get the textual embedding for the Emotion Recognition task. One notable BERT variation is SapBERT [4] which incorporates self-align embedding in its architecture. SapBERT introduces a custom loss function to align the representation space of biomedical entities, making it a state-of-the-art model in the medical domain. The use of self-align can be broadened to any field that requires adjusting feature vectors.

### C. Multimodal speech emotion recognition

In the field of SER, recent research has made significant advancements. In 2023, a multimodal approach called SERVER [2] combines the audio feature and text feature using BERT and VGGish-based respectively. This fusion of audio and text features allows the model to gain deeper insights from multiple input data, leading to improved performance. SERVER demonstrated its competitiveness by outperforming many previous methods in multimodal SER. However, SERVER had a limitation as it simply concatenated the audio and text features without considering their interdependencies. To address this issue, a subsequent method called 3M-SER [3] was proposed. 3M-SER introduced a fusion module that employs a self-attention mechanism [9] to analyze the relationships between audio and text features. By capturing the relevant connections between these modalities, 3M-SER achieved remarkable performance compared to previous approaches in multimodal SER. Another approach in SER is MMER [10] which uses multi-head attention fusion and multi-feature embeddings. MMER introduces the crossmodal encoder which can achieve over 80% of both weighted and unweighted accuracy when evaluating on IEMOCAP [11] dataset.

## III. DATASETS

### A. IEMOCAP

The Interactive Emotional Dynamic Motion Capture (IEMOCAP) [11] dataset, introduced in 2008, is a valuable resource for studying speech emotions. It is available in the English language and was recorded at the University of Southern California. The corpus features were recorded by 10 professional actors and divided into five separate sessions. Each session involves both a male and a female actor. The IEMOCAP dataset consists of audio-visual files, with each file spanning approximately 12 hours in length. The recorded

utterances, on average, have a duration of around 3.5 seconds. The dataset covers a range of different emotions, making it suitable for various emotion recognition studies.

TABLE I: The contributions of six major emotions in the IEMOCAP dataset.

Emotion	Number of samples	Distribution
Neutral	1849	25.05%
Frustrated	1708	23.14%
Sad	1103	14.95%
Anger	1084	14.69%
Excited	1041	14.11%
Happy	595	8.06%

In this study, we evaluate our method on two scenarios: a 4-class and a 6-class emotions classification task, utilizing the emotions present in the IEMOCAP dataset. Table I gives details of the speech emotions, audio file quantity, and contribution rate of each emotion which are used in this study. In our evaluation, we focused on four major classes, namely neutral, sad, anger, and happy, to assess the effectiveness of our method for the 4-class emotions classification task, in which happy and excited is grouped together as happy. For the 6-class emotion classification task, we utilized all the emotions listed in Table I to train and evaluate our model.

### B. MELD

The Multimodal EmotionLines Dataset (MELD) [12] is an expanded and enriched version of the EmotionLines [13] dataset. MELD includes the same conversational exchanges found in EmotionLines, but it incorporates multiple modalities, encompassing audio, visual, and text data. MELD comprises over 1,400 dialogues and 13,000 individual utterances extracted from TV series, featuring contributions from various speakers. Each utterance within a conversation has been assigned one of seven distinct emotional labels, which include Anger, Disgust, Sadness, Joy, Neutral, Surprise, and Fear. The distribution of these emotion categories can be found in Table II. The same approach of selecting emotions for the 4-class evaluations is also applied to the MELD dataset. To provide consistency across two datasets, we map emotions in MELD accordingly to emotions in IEMOCAP, namely joy is mapped to happy, sadness to sad, and anger to anger. As a result, the 4-class emotion for MELD is now neutral, sad, anger, and happy, which is similar to IEMOCAP.

TABLE II: The contributions of six major emotions in the MELD dataset.

Emotion	Number of Samples	Distribution
Neutral	4710	47.15%
Joy	1743	17.45%
Surprise	1205	12.06%
Anger	1109	11.10%
Sadness	683	6.84%
Disgust	271	2.71%
Fear	268	2.68%

## IV. PROPOSED METHOD

To help the model gain insight into feature representation in latent space, we have developed a model called MERSA which combines information from both audio and text inputs to effectively tackle the emotion recognition task. To extract features for the text input, we have created our custom pre-trained model called self-align BERT embedding. This model captures important textual information related to emotions by combining BERT [8] with self-align embedding [4]. For the audio input, we utilize either VGGish or Wav2Vec feature extractors to obtain acoustic feature vectors. These feature vectors capture relevant acoustic information associated with emotions. To integrate the information from both modalities, we employ a cross-attention fusion module. This module enables us to capture rich and meaningful information between the text and audio inputs, enhancing the overall emotion recognition capability of the model. Fig. 1 provides an overview of the working process of the MERSA model, illustrating how the different components interact and contribute to the model’s performance.

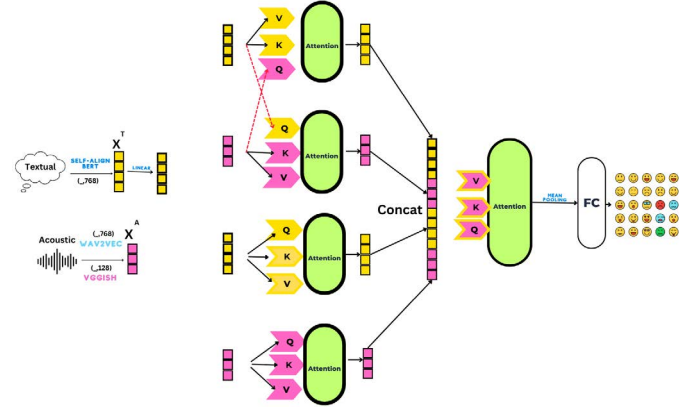


Fig. 1: The proposed flowchart of MERSA model.

### A. Self-Align Embedding

Inspired by SapBERT [4], we designed a framework to self-align textual feature vectors using BERT-based [8] as the foundation. The self-alignment method is a learning method geared towards learning effective feature vectors (or representations) of emotion data in a way that captures the underlying structure and relationships among the samples and the emotions. Our research applies Multi-similarity loss [14], which is a concept derived from metric learning. The Multi-similarity loss technique focuses on minimizing intra-class variances while maximizing inter-class margins, which enables the model to self-align textual feature vectors as a pre-processing method for the proposed MERSA model. In the context of feature vectors, it involves adjusting the distances between feature vectors such that similar vectors are brought closer, and dissimilar ones are pushed apart in the latent space. This loss function is particularly effective in scenarios where the relationships within the data are complex and multi-faceted.

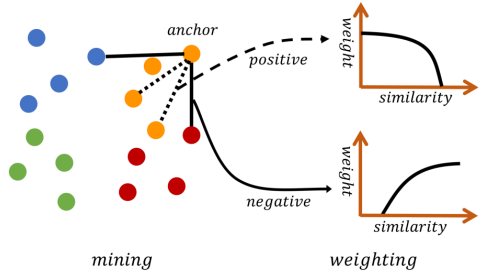


Fig. 2: Illustration of multi-similarity loss [14]

Fig. 2 visualizes the concept of the Multi-similarity loss function ( $L_{MS}$ ) which has the formula as follows:

$$L_{MS} = \frac{1}{m} \sum_{i=1}^m \left[ \frac{1}{\alpha} \log \left( 1 + \sum_{k \in P_i} e^{-\alpha(S_{ik}-\lambda)} \right) + \frac{1}{\beta} \log \left( 1 + \sum_{k \in N_i} e^{\beta(S_{ik}-\lambda)} \right) \right] \quad (1)$$

where  $m$  is the number of training samples or pairs in the batch, used to compute the average loss,  $P_i$  represents a set of positive pairs for the  $i^{th}$  sample, pairs of data points that are similar or related,  $N_i$  represents a set of negative pairs for the  $i^{th}$  sample, consisting of dissimilar or unrelated data points.  $S_{ik}$  is the similarity score between the  $i^{th}$  and  $k^{th}$  samples, measuring their similarity.  $\lambda$  is the margin parameter defining the boundary between positive and negative pairs,  $\alpha$  and  $\beta$  are the scaling parameters for the exponential functions, controlling the change in pair contributions to the loss with their relative similarities.

This technique offers a sophisticated means to enhance the representational capacity of models, ensuring that the feature vectors accurately reflect the inherent similarities and differences within the data. The terms  $e^{-\alpha(S_{ik}-\lambda)}$  and  $e^{\beta(S_{ik}-\lambda)}$  are used to compute the weights for positive and negative pairs, respectively. The exponential function ensures that pairs with higher similarity (for positive pairs) or lower similarity (for negative pairs) relative to the margin  $\lambda$  are given more weight. The logarithm is applied to the sum of the exponential terms to stabilize the training and prevent numerical issues due to the potentially large range of the exponential function. The loss is computed by summing over all the training samples in the mini-batch and then averaging the sum.

### B. Self-Attention and Cross-Attention Module

Our self-attention module uses a multi-head attention [9] block to determine which features in the text and audio feature vectors are relevant to the classification process. This block also aligns the dimensions of text feature vectors with audio feature vectors through a linear transformation and LayerNorm [15]. A multi-head attention module is composed of multiple single-head attention modules that are stacked together. Each head operates independently but in parallel, allowing the model to capture different aspects of the input and

enhance its representational capacity. The formula of single-head attention is as follows:

$$\begin{aligned} Head_{self} &= Attention(Q_{self}, K_{self}, V_{self}) \\ &= Softmax\left(\frac{Q_{self}K_{self}^T}{\sqrt{d_k}}\right)V_{self} \end{aligned} \quad (2)$$

where  $Q_{self}$ ,  $K_{self}$ , and  $V_{self}$  represent the query, key, and value respectively in a single-head, and  $d_k$  represents the feature dimension of the key. The attention mechanism calculates  $Head_{self}$  which is a weighted sum of the values  $V_{self}$  based on the similarity between the query  $Q_{self}$  and the key  $K_{self}$ . The weighted sum is parameterized by the attention weights obtained from applying the softmax function on the similarity by the dot product of  $Q_{self}$  and  $K_{self}$  divided by the square root of  $d_k$ .

Since the original feature values of text and audio may have significant differences due to their respective extraction models, we apply a LayerNorm [15] to both the audio and text feature vectors. This normalization brings the values of text and audio closer together, ensuring fairness in their contribution to the final model. This research also incorporates and experiments with an advanced cross-attention module into the MERSA model as shown in Fig. 1. The proposed cross-attention architecture diverges from the traditional self-attention mechanism, predominantly by varying the input feature vectors for the query ( $Q$ ), key ( $K$ ), and value ( $V$ ) components. In this architecture, the text feature vectors are utilized as inputs for the query component, while the audio feature vectors are employed for both the key and value components. This process is then inverted, wherein the audio feature vectors are used for the query, and text feature vectors are utilized for the key and value. Such an approach allows for a more intricate interplay between the modalities and potentially enhances the recognition accuracy of complex emotional states. The formula for the text-audio cross-attention  $CrossAtt_{text-audio}$  is as follows:

$$CrossAtt_{text-audio} = Attention(Q_{text}, K_{audio}, V_{audio}) \quad (3)$$

And the formula for the audio-text cross-attention  $CrossAtt_{audio-text}$  is as follows:

$$CrossAtt_{audio-text} = Attention(Q_{audio}, K_{text}, V_{text}) \quad (4)$$

We combine both crossmodel attention and self-attention architecture as Fig. 1. While the self-attention module helps each modality capture the intra-connection information in itself, the cross-attention mechanism helps capture the relationships between textual and audio information, allowing the model to attend to relevant features from both modalities in a coordinated approach.

## V. EXPERIMENTAL RESULT AND DISCUSSION

### A. Experimental Setup

Our research focuses on two primary datasets: MELD and IEMOCAP. For the MELD dataset, we train our models to

classify emotions into 4-class categories. Similarly, for the IEMOCAP dataset, we conduct training for emotion classification into 4-class and 6-class categories. We begin by evaluating a baseline model to establish a foundational performance metric. Following this, we incrementally introduce additional modules to observe their respective contributions to the model’s effectiveness. The progression of our experimental setup is as follows: First, we introduce a baseline model inspired by SERVER [2], then integrate the Multi-head Attention module to a multimodal model similar to the approach shown in 3M-SER [3]. Next, we introduced our self-align technique textual feature vectors. The audio feature vectors method is tested with both VGGish and Wav2Vec. Our final model includes the incorporation of the Cross Modality Attention Module, which is expected to effectively synthesize data from both textual and acoustic modality better. Regarding our training model, we have opted for a batch size of 1. This decision is primarily driven by the variable length of audio feature vectors corresponding to each sentence in a conversation. Utilizing a uniform batch size would necessitate padding, which could potentially reduce the model’s accuracy due to the non-informative data. By adopting a batch size of 1, we ensure that each input is processed in its original form, thereby maintaining the integrity of the data and enhancing the model’s performance. The other settings follow the original settings in SERVER and 3M-SER to train and evaluate the model.

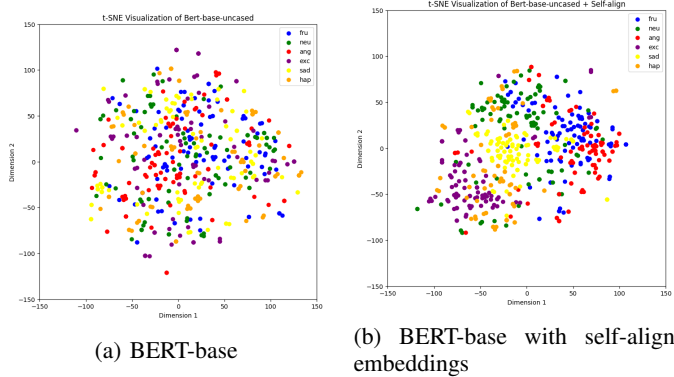


Fig. 3: T-SNE plot of BERT-base and BERT-base with self-align embeddings

### B. Self-align embeddings

To improve the performance of BERT-base [8] extractor, we integrate self-align textual embeddings into BERT-base architecture. Fig. 3 visualizes the T-SNE plot of BERT-base and BERT-base with self-align embeddings, which shows significant improvement in terms of sample distribution and representation. Although similar vector features are closer together, there is still overlap between different emotion categories. This overlap occurs when we visualize high-dimensional data using dimensionality reduction methods, some information may be lost. Nevertheless, based on T-SNE visualization, integrating self-align textual embeddings into the BERT-base enables the

model to generate better feature vectors in the latent space. With this approach, the model can converse faster as well as improve the performance of the fusion module when used in the multimodal architecture.

### C. Experiment results

To compare different models, we focused on training the model using only four main classes as mentioned in dataset section: anger, happy, sad, and neutral. We divided our dataset into the train, validation, and test sets using a ratio of 8:1:1. During the training phase, we applied the early stopping strategy to prevent overfitting on the training dataset.

TABLE III: Performance comparison of models training on IEMOCAP (4 emotions)

Model	Method	Accuracy	F1
SERVER [2]	VGGish and BERT	0.616	0.612
SERVER modified	Wav2Vec and BERT	0.604	0.622
3M-SER [3]	VGGish and BERT	0.751	0.745
3M-SER modified	Wav2Vec and BERT	0.777	0.784
MERSA (Ours)	VGGish and BERT self-align	0.770	0.764
MERSA (Ours)	VGGish and BERT self-align with cross-attention	<b>0.805</b>	0.791
MERSA (Ours)	Wav2Vec and BERT self-align	0.783	<b>0.799</b>
MERSA (Ours)	Wav2Vec and BERT self-align with cross-attention	0.779	0.764

In our experiments, we evaluated two common models that are utilized for audio processing: VGGish [6] and Wav2Vec [7]. To compare the MERSA model with SERVER [2] and 3M-SER [3] using Wav2Vec, we replaced the VGGish backbone in both models with Wav2Vec, resulting in a modified version of the architecture as shown in Table III, which presents the performance comparison of various models on the IEMOCAP dataset. In which both cross-attention and self-align show a performance improvement compared to the previous methods. Specifically, the MERSA model, incorporating self-align, achieves the highest scores with an accuracy of 0.805 and an F1 score of 0.799. This once again demonstrates the effectiveness of using self-align embeddings to enhance the model’s performance on the SER tasks. Additionally, when cross-attention is applied to the MERSA model with VGGish as an audio extractor, there is an improvement in the model’s accuracy and F1 score compared to previous methods. However, interestingly, when we use Wav2Vec as the audio extractor instead, the performance of the MERSA model becomes worse than not using any cross-attention at all. The reasons behind this discrepancy can be related to limited discriminative information. The features extracted by Wav2Vec may not capture sufficient discriminative information for the specific task being performed. The representations learned by VGGish may be more suitable and task-specific, leading to better performance when used in conjunction with the cross-attention mechanism. In our experiments, we only focus on fine-tuning the textual model and using the pre-trained audio processing models without considering its feature vectors in latent space. In the future, the problem related to audio feature vectors will be considered to further improve the performance of multimodal models.

TABLE IV: Performance comparison of models training on IEMOCAP (6 emotions)

Model	Method	Accuracy	F1
SERVER [2]	VGGish and BERT	0.535	0.478
SERVER modified	Wav2Vec and BERT	0.521	0.502
3M-SER [3]	VGGish and BERT	0.586	0.577
3M-SER modified	Wav2Vec and BERT	0.595	0.584
MERSA (Ours)	VGGish and BERT self-align	0.644	0.621
MERSA (Ours)	VGGish and BERT self-align with cross-attention	0.669	0.653
MERSA (Ours)	Wav2Vec and BERT self-align	0.671	<b>0.655</b>
MERSA (Ours)	Wav2Vec and BERT self-align with cross-attention	<b>0.676</b>	0.646

We also trained and evaluated it on the 6-class IEMOCAP dataset, as displayed in Table IV. Once again, the MERSA model with self-align embeddings achieved the highest scores among all methods in terms of F1 score even with the increase in the number of classes. While MERSA model with a cross-attention model performs slightly better in accuracy compared to using only self-attention module, this outcome still strongly affirms that both self-align embeddings and cross-attention can effectively enhance the feature vectors for multimodal models.

TABLE V: Performance comparison of models training on MELD (4 emotions)

Model	Method	Accuracy	F1
SERVER [2]	VGGish and BERT	0.562	0.424
SERVER modified	Wav2Vec and BERT	0.5851	0.592
3M-SER [3]	VGGish and BERT	0.637	0.509
3M-SER modified	Wav2Vec and BERT	0.653	0.627
MERSA (Ours)	VGGish and BERT self-align	0.665	0.602
MERSA (Ours)	VGGish and BERT self-align with cross-attention	0.661	0.612
MERSA (Ours)	Wav2Vec and BERT self-align	<b>0.677</b>	<b>0.634</b>
MERSA (Ours)	Wav2Vec and BERT self-align with cross-attention	0.672	0.641

To further evaluate the performance of the MERSA model, we tested on the MELD dataset and the results are shown in Table V, we present the performance comparison between our method and previous studies on the MELD dataset. Once again, our MERSA model shows an improvement over previous architectures and achieved the highest performance with an accuracy of 0.677 and an F1 score of 0.634.

## VI. CONCLUSION

In this paper, a multimodal architecture has been proposed to improve the performance of multimodal SER named MERSA. MERSA model enhances the fusion feature vectors of text and audio by leveraging the self-align embeddings and cross-attention module. The experimental results have shown that the MERSA model improved the performance of the previous multimodal model, which achieved the highest scores compared to the previous study. The implementation of the self-align technique consistently improved upon existing models. It yielded a substantial increase in both accuracy and F1 score. However, this method also presented challenges, such as the potential for overfitting in certain high-performance models and a reduction in inference time. The addition of

the cross-attention module shows mixed results, the minor improvement in benchmark score does not seem to justify the increase in computing resource cost. This outcome suggests the need for further optimization in audio feature vectors to incorporate the cross-attention module into MERSA.

## REFERENCES

- [1] J. Li, Y. Liu, X. Wang, and Z. Zeng, "Cfn-esa: A cross-modal fusion network with emotion-shift awareness for dialogue emotion recognition," *arXiv preprint arXiv:2307.15432*, 2023. [Online]. Available: <http://arxiv.org/abs/2307.15432>
- [2] N. T. Pham, D. N. M. Dang, B. N. H. Pham, and S. D. Nguyen, "Server: Multi-modal speech emotion recognition using transformer-based and vision-based embeddings," in *ACM International Conference Proceeding Series*, 2023, pp. 234–238.
- [3] P.-N. Tran, T.-D. T. Vu, D. N. M. Dang, N. T. Pham, and A.-K. Tran, "Multi-modal speech emotion recognition: Improving accuracy through fusion of vggish and bert features with multi-head attention," in *Industrial Networks and Intelligent Systems*, N.-S. Vo and H.-A. Tran, Eds. Cham: Springer Nature Switzerland, 2023, pp. 148–158.
- [4] F. Liu, E. Shareghi, Z. Meng, M. Basaldella, and N. Collier, "Self-alignment pretraining for biomedical entity representations," *arXiv preprint arXiv:2010.11784*, 2020. [Online]. Available: <http://arxiv.org/abs/2010.11784>
- [5] C.-F. R. Chen, Q. Fan, and R. Panda, "Crossvit: Cross-attention multi-scale vision transformer for image classification," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 347–356.
- [6] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "Cnn architectures for large-scale audio classification," *arXiv preprint arXiv:1609.09430*, 2016. [Online]. Available: <http://arxiv.org/abs/1609.09430>
- [7] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," *arXiv preprint arXiv:1904.05862*, 2019. [Online]. Available: <http://arxiv.org/abs/1904.05862>
- [8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018. [Online]. Available: <http://arxiv.org/abs/1810.04805>
- [9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017. [Online]. Available: <http://arxiv.org/abs/1706.03762>
- [10] S. Ghosh, U. Tyagi, S. Ramaneswaran, H. Srivastava, and D. Manocha, "Mmer: Multimodal multi-task learning for speech emotion recognition," 2023.
- [11] C. Busso, M. Bulut, C. C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, 2008. [Online]. Available: <https://doi.org/10.1007/s10579-008-9076-6>
- [12] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," *arXiv preprint arXiv:1810.02508*, 2018. [Online]. Available: <http://arxiv.org/abs/1810.02508>
- [13] C.-C. Hsu, S.-Y. Chen, C.-C. Kuo, T.-H. Huang, and L.-W. Ku, "EmotionLines: An emotion corpus of multi-party conversations," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, and T. Tokunaga, Eds. Miyazaki, Japan: European Language Resources Association (ELRA), May 2018.
- [14] X. Wang, X. Han, W. Huang, D. Dong, and M. R. Scott, "Multi-similarity loss with general pair weighting for deep metric learning," *CoRR*, vol. abs/1904.06627, 2019. [Online]. Available: <http://arxiv.org/abs/1904.06627>
- [15] J. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *ArXiv*, vol. abs/1607.06450, 2016. [Online]. Available: <https://api.semanticscholar.org/CorpusID:8236317>