



Modality-Agnostic Multimodal Emotion Recognition using a Contrastive Masked Autoencoder

Georgios Chochlak¹, Turab Iqbal², Woo Hyun Kang², Zhaocheng Huang²

¹University of Southern California, USA

²AWS AI Labs, USA

chochlak@usc.edu, {iqbturab,whkang,davidhzc}@amazon.com

Abstract

Multimodal deep learning methods have greatly accelerated research in emotion recognition and have become the state of the art. However, in many scenarios, not all modalities are readily available, leading to either failure of traditional algorithms or the need for multiple models. In this work, we advance the state of the art in emotion recognition by proposing a unified, modality-agnostic transformer-based model that is inherently robust to missing modalities. To better exploit the multimodality of the data, we propose to use contrastive learning for modality alignment and masked autoencoding for multimodal reconstruction. Experimental results on the MSP-Podcast corpus show that our unified model achieves state-of-the-art performance, and improves both unimodal and multimodal baselines by 1-5% relative in respective evaluation metrics with the capability to handle missing modalities for two emotion recognition tasks in a more compact model.

Index Terms: Multimodal Emotion Recognition, Missing Modality

1. Introduction

The automatic detection of human emotions from behavioral signals (e.g., voice) plays a crucial role in enabling natural, engaging interactions between humans and machines, and has gained increasing popularity within the community over the past two decades. Due to the complementarity among modalities, multimodal systems have been widely adopted for improved performance and robustness of emotion recognition systems. Multimodal emotion recognition has greatly benefited from the recent proliferation of deep learning, ranging from emotionally informative representations learned from various modalities such as language [1, 2], audio [2, 3, 4], vision [3, 4], or physiological signals, to generalized downstream models for a plethora of domains and use cases.

Nonetheless, in realistic scenarios, multimodal systems often necessitate all input modalities, which limits their ability to robustly handle cases where one modality is missing (e.g., due to malfunctioning/disabled microphones or cameras in a virtual meeting). For example, an audio+text model may not necessarily handle text-only social media posts as well as audio recordings with text transcriptions. This means that multiple models are required in practice to handle scenarios of different possible missing modalities.

Within the research community, there has been growing interests in resolving the complete missing modality issue. Some studies investigated joint training of unimodal and multimodal downstream heads trained on top of a shared network [3, 5]. A number of studies handled the missing modality scenario via generating or reconstructing pseudo embeddings for the

missing modalities from the available modalities, and combine them for emotion classification [6, 7, 8, 9, 10, 11]. However, existing methods require additional convoluted reconstruction/generation steps for the missing modalities at the utterance level with extra networks/computation and loss of more granular information.

In this work, we propose a unified model to handle completely missing modalities via a shared transformer operating on concatenated *sequences* of embeddings from different modalities. The shared transformer is modality-agnostic, as it operates on any available modality or combination thereof. We introduce two additional objectives: a contrastive objective for aligning the modalities at the utterance level, and a multimodal masked autoencoder (MAE) to learn multimodal interactions at the frame or token level. The proposed framework is effective in removing extra networks and leveraging more detailed information for reconstruction.

Our study has two main contributions: 1) We take a different approach to emotion recognition under missing modalities by creating a unified model that is inherently robust to the issue. 2) We enhance the cross-modal knowledge of our model with two self-supervised multimodal tasks, MAE and contrastive learning, enabling reconstruction and modality alignment/retrieval in addition to emotion recognition.

2. Related Work

The simplest approach to deal with missing modalities would be to jointly train the unimodal and multimodal systems with a shared backbone network, such as proposed in [3, 5], which utilizes unimodal classifiers when only one modality is present, and a multimodal classifier when all modalities are present. Such a framework, however, cannot explore the rich multimodal structure from the emotional data and requires multiple models to be hosted in production.

Other previous works have relied on the conditional generation of missing modalities given the present ones [6, 7, 9]. More specifically, an encoder was adopted per modality (e.g., HuBERT [12] for audio, RoBERTa [13] for text) for generating embeddings used in downstream classifiers, and embeddings for the missing modalities were conditionally generated from the available modality embeddings. The generation or reconstruction can be done via a cascade of autoencoders [6], a diffusion model [7], or a transformer-based module [8, 9], given the available modality embeddings. Additional advancements have also been made along this path. For example, CIF-MMIN [10] and IF-MMIN [11] highlight the advantage of learning modality-invariant representations prior to the reconstruction step, with CIF-MMIN using contrastive learning and IF-MMIN employing central moment discrepancy distance. Recently, Lian et al

[14] introduced Graph Complete Network (GCN) to deal with the missing modality issue in conversations. Such setup allowed the reconstruction to benefit from semantically related adjacent utterances in conversations.

Although these studies can address the missing modality problem, there remains limitations such as loss of fine-grained information, overly convoluted model architectures, or the need for multi-stage training [10]. More specifically, these methods rely on utterance-level embeddings for reconstruction, which might not sufficiently capture within-utterance multimodal patterns or interactions. Moreover, the embedding generation or reconstruction for the missing modality often require additional explicit, convoluted networks (one per each missing modality) while the complementarity of the reconstructed embeddings remains relatively questionable to the downstream tasks. In order to overcome these limitations, we leverage a more fine-grained representation at the frame or token level, rather than at the utterance level. Moreover, we enabled the model to have built-in reconstruction capabilities without the explicit embedding generation.

Furthermore, motivated by recent advances of self-supervised multimodal models that effectively capture multimodal interactions, we hypothesize that such cross-modal information may benefit the robustness to the missing modality issue. A seminal work is CLIP [15], where a contrastive objective was proposed to align an image encoder with a language model for image captioning. This was expanded to align different modalities [16, 17] or integrate other objectives like masking [17] and denoising [18]. Thus, we explore whether the enhanced modality alignment can aid token-level reconstruction for emotion recognition with missing modalities. It is worth noting that such contrastive objective is different from [10], where contrastive loss was used to learn modality-invariant features.

In this study, we focus solely on the audio and text modalities, driven by the fact that visual signals are not always available, especially in real-world applications such as contact centers and podcasts, where emotion recognition tends to be the primary use case.

3. Proposed Framework

3.1. Modality-agnostic Shared Transformer

As illustrated in Fig. 1, given the raw audio and text inputs for each sample i of the labeled dataset $D = \{(\mathbf{x}_i^a, \mathbf{x}_i^t, y_i)\}_i$, we use a pre-trained encoder, $E^m, m \in \{a, t\}$, to extract a sequence of embeddings (including all intermediate layers) for each modality $\{\mathbf{H}_{i,j}^{m,l}\}_{l,j} = E^m(\mathbf{x}_i^m)$, and then we derive a weighted embedding by computing a convex combination of the layers at the j -th frame or token:

$$\mathbf{H}_{i,j}^m = \sum_l w_l^m \mathbf{H}_{i,j}^{m,l} \quad (1)$$

where $w_l^m \geq 0$ are learnable weights, l is the layer index, and $\sum_l w_l^m = 1$. Then, we project both sequences to the input space of the shared transformer. Since audio sequence lengths are much greater than text ones, we downsample the audio sequence at the time dimension using a convolutional layer with a stride equal to the kernel size k , as opposed to the linear layer for text:

$$\begin{aligned} \mathbf{h}_{i,j'}^a &= \text{Conv}(\{\mathbf{H}_{i,j}^a\}_{j=j' \cdot k}^{(j'+1) \cdot k}) \\ \mathbf{h}_{i,j}^t &= \text{Linear}(\mathbf{H}_{i,j}^t) \end{aligned} \quad (2)$$

The resulting sequences were concatenated alongside two special tokens, CLS and SEP [19]. CLS is prepended and used for supervised tasks, such as classification, while SEP is interjected between the two modalities. Further, we add modality-type embeddings \mathbf{v}^m to each modality, as in [20]:

$$\mathbf{h}_i^{at} = \{\mathbf{h}^{CLS} | \{\mathbf{h}_{i,j}^a + \mathbf{v}^a\}_{j=1}^{n_i^a} | \mathbf{h}^{SEP} | \{\mathbf{h}_{i,j}^t + \mathbf{v}^t\}_{j=1}^{n_i^t}\} \quad (3)$$

where $|$ denotes concatenation. When a modality is missing, we simply skip its embedding sequence in the concatenation. Three separate passes were generated to go through the shared transformer, one for each combination of modalities.

The shared transformer (ST) generates contextual embeddings $\mathbf{z}_i = ST(\mathbf{h}_i^{at})$ for each pass. We use \mathbf{z}_i^{CLS} to perform supervised tasks, such as emotion classification for emotion categories, or regression for emotion dimensions, with the corresponding supervised loss, $\mathcal{L}_S(f(\mathbf{z}_i^{CLS}), y_i)$, where f is a two-layer feedforward network.

3.2. Masked Autoencoder (MAE)

In addition, we propose to mask tokens at the input of the shared transformer with an additional embedding, MASK [19], and teach the model to reconstruct them with the mean squared error loss, effectively training the shared transformer to be a MAE. We found that it is more effective to reconstruct the encoder embeddings \mathbf{H} than the input to the shared transformer \mathbf{h} , analyzed in Section 4.4. To do so, we need to reverse the projection (and the temporal downsampling). For text, a linear layer suffices, whereas for audio, we use a transposed convolution layer:

$$\begin{aligned} \hat{\mathbf{H}}_{i,j}^t &= \text{Linear}(\mathbf{z}_{i,j}^t) \\ \{\hat{\mathbf{H}}_{i,j}^a\}_{j=j' \cdot k}^{(j'+1) \cdot k} &= \text{ConvT}(\mathbf{z}_{i,j'}^a) \end{aligned} \quad (4)$$

Finally, we set

$$\mathcal{L}_{MAE} = \frac{1}{N} \sum_m \sum_i \frac{1}{n_i^m} \sum_j \frac{1}{d^m} \|\hat{\mathbf{H}}_{i,j}^m - \mathbf{H}_{i,j}^m\|^2 \quad (5)$$

where N is the batch size, and d^m the dimension of \mathbf{H}^m .

3.3. Contrastive Learning

Motivated by CLIP [15], we propose to enhance the inter-modality alignment with contrastive learning. For each of the unimodal runs, we extract a modality embedding per utterance from the contextual embeddings with mean pooling, and then passing them through a feedforward network g , i.e., $\mathbf{z}_i^m = g(\frac{1}{n_i^m} \sum_j \mathbf{z}_{i,j}^m)$. We then compute the contrastive loss within a batch of N utterances as the semisum of the cross-entropy losses across columns and rows of the similarity matrix:

$$\mathcal{L}_C = -\frac{1}{2N} \sum_i \left(\log \frac{\mathbf{z}_i^a \cdot (\mathbf{z}_i^t)^T}{\sum_{i'} \mathbf{z}_{i'}^a \cdot (\mathbf{z}_{i'}^t)^T} + \log \frac{\mathbf{z}_i^t \cdot (\mathbf{z}_i^a)^T}{\sum_{i'} \mathbf{z}_{i'}^t \cdot (\mathbf{z}_{i'}^a)^T} \right) \quad (6)$$

g is a linear layer plus non-linearity, which helps avoid overfitting the modality embeddings to the contrastive task [21]. We expect that the alignment between modalities will ground one to the other, creating text-informed audio representations and vice versa, thus improving unimodal performance. Since the shared transformer extracts meaningful embeddings from only one modality, this can further improve unimodal performance.

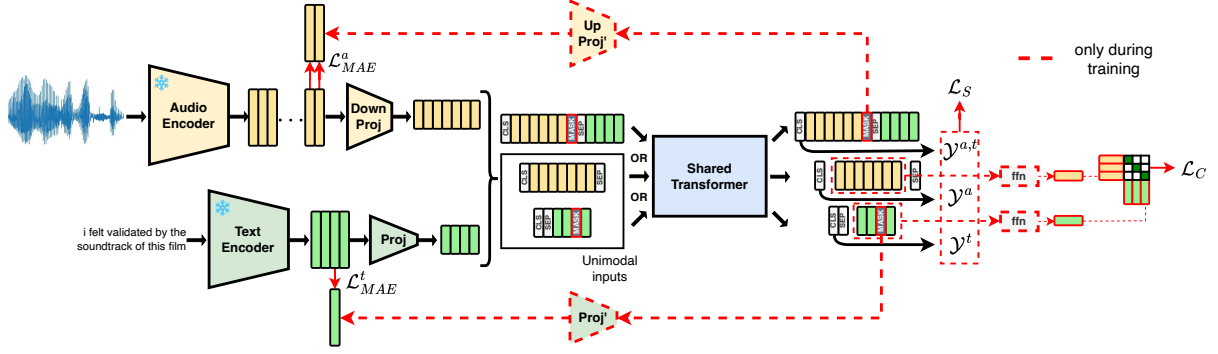


Figure 1: Our proposed framework: After extracting embeddings from frozen encoders, we project them to the input space of the shared transformer, followed by a simple concatenation of sequential "tokens" (Eq. (3)). The audio-only, text-only and audio + text sequences are processed through the same model. The CLS token is trained with supervision using \mathcal{L}_S . We mask encoder embeddings \mathbf{H}^m and learn to reconstruct them with \mathcal{L}_{MAE} using the shared transformer. Finally, we create a multimodal contrastive objective, \mathcal{L}_C , within the batch for modality alignment.

3.4. Entire Framework

Summing up, the overall loss for training the shared transformer is, as shown in Fig. 1:

$$\mathcal{L} = \mathcal{L}_S + \lambda_{MAE} \mathcal{L}_{MAE} + \lambda_C \mathcal{L}_C, \quad (7)$$

where λ are hyperparameters. Note that, for each example, we perform separate forward passes for the MAE in order to avoid contamination from the masks in classification and contrastive learning, though only one backpropagation pass. As a result, the model can perform emotion recognition, reconstruction, and audio and text modality alignment. Finally, as shown in Fig. 1, the components required for the MAE and contrastive learning can be discarded during inference.

4. Experiments

4.1. Datasets

We used the MSP-Podcast corpus (version 1.11) [27], which is among the largest and widely adopted publicly available emotional speech corpora in English, containing 238 hours of naturalistic speech. MSP-Podcast significantly outsize existing commonly adopted emotional datasets such as IEMOCAP, CMU-MOSEI, CMU-MOSI, MSP-IMPROV, etc and hence is preferred. We used the default partition that has 84030, 19815, and 30647 utterances for training, validation and evaluation, respectively, where each utterance was annotated in terms of categorical (e.g., neutral, anger) and dimensional (e.g., arousal) emotions at the utterance level from multiple annotators. In this study, we address two emotion tasks: emotion classification and prediction. For emotion classification, we adopted the proportion of the dataset that has majority consensus on four main emotion categories, namely neutral, happiness, sadness, and anger. For prediction of emotion dimensions, we adopted the entire dataset, and use the mean scores among all annotators. We also converted the valence score into neutral, positive, and negative sentiment with a threshold of 3 and 5, and added a task for sentiment classification to aid overall performance.

4.2. Implementation Details

In MSP-Podcast, we set \mathcal{L}_S to be the average of all the sentiment cross entropies, the emotion cross entropies, and the losses

based on Concordance Correlation Coefficients (CCC) for all three emotion dimensions. We empirically set $\lambda_{MAE} = 10$ and $\lambda_C = 1$. For the shared transformer, we use RoPE [28], train with 50% dropout, set the hidden dimension to 128, and search for the number of layers in $\{2, 4, 8\}$. The audio sequences were downsampled by a factor of $k = 4$. For the contrastive loss, we do not use a temperature, nor normalize the dot products, and use the ReLU in g . For MAE, we use a 50% masking rate, and found the transposed convolution to work better than the custom deconvolution for reconstruction.

We evaluated the systems using Unweighted Accuracy (UA) for emotion classification and CCC for emotion prediction [2]. We used RoBERTa-large as the text encoder and Whisper-medium [29] as the audio encoder, which we found to work better than HuBERT-large and WavLM-large [30]. Standard deviations across 3 runs were $< 0.2\%$ for accuracies and < 0.005 for CCCs.

4.3. Effectiveness of the Shared Transformer

For benchmarking against the state-of-the-art in handling missing modalities, we implemented the CIF-MMIN system in [10] using the open-source code. The code was adapted to handle the audio and text modalities with the same pre-trained encoders adopted in this study, i.e., RoBERTa-large and Whisper-medium. Moreover, we implemented three additional baselines: 1) **Cascade**: We train three separate classifiers concurrently on top of both unimodal and multimodal CLS embeddings. 2) **Separate models**: We train three separate models, one for each possible combination of modalities; 3) **Late fusion**: We train two unimodal models, and then use these pretrained models to train a late fusion classifier on top of their concatenated CLS embeddings;

Table 1 summarizes results of the proposed framework in comparison to the literature and the implemented baseline systems. There are three major observations.

First, the proposed shared transformer (w/ or w/o contr+MAE) achieve state-of-the-art performance for both emotion classification and prediction in both unimodal and multimodal scenarios using a single model.

Second, the proposed shared transformer outperforms our baseline implementations, including CIF-MMIN that was proposed to deal with the missing modality issue. It is worth not-

Table 1: Comparison of our proposed approach to baselines in terms of UA and CCCs for all modality combinations. Note that the proposed shared transformer is a unified model for all the tasks and input modality combinations. The first five rows show state-of-the-art results reported in the literature. We implemented the CIF-MMIN system with adaptation to the audio and text modalities.

Model	UA (%)			Valence CCC			Arousal CCC			Dominance CCC		
	<i>a</i>	<i>t</i>	<i>a, t</i>	<i>a</i>	<i>t</i>	<i>a, t</i>	<i>a</i>	<i>t</i>	<i>a, t</i>	<i>a</i>	<i>t</i>	<i>a, t</i>
SER-ASR (2024) [22]	-	38.2	61.6	-	0.526	-	-	0.316	-	-	0.293	-
SpeechVerse (2024) [23]	65.1	-	-	-	-	-	-	-	-	-	-	-
KNN-VC (2024) [24]	-	-	-	0.568	-	-	0.656	-	-	0.485	-	-
Multilingual-LoRA (2024) [25]	-	-	-	0.647	-	-	0.680	-	-	0.616	-	-
wavLM-LR (2024) [26]	-	-	-	0.655	-	-	0.679	-	-	0.628	-	-
CIF-MMIN (2024) [10]	56.9	44.0	59.6	-	-	-	-	-	-	-	-	-
Cascade	58.3	48.5	60.8	0.650	0.565	0.688	0.673	0.316	0.676	0.611	0.310	0.616
Separate models	65.6	54.3	66.8	0.645	0.545	0.683	0.673	0.324	0.676	0.603	0.301	0.607
Late fusion	65.6	54.3	61.0	0.645	0.545	0.675	0.673	0.324	0.671	0.603	0.301	0.598
Shared Transformer	64.2	54.4	67.6	0.656	0.568	0.701	0.683	0.328	0.682	0.612	0.317	0.616
+contr	65.1	55.0	67.9	0.642	0.568	0.694	0.678	0.329	0.679	0.613	0.310	0.614
+MAE	63.2	54.6	67.3	0.632	0.573	0.700	0.680	0.334	0.673	0.607	0.318	0.602
+contr+MAE	66.6	54.5	68.3	0.656	0.571	0.694	0.679	0.331	0.677	0.612	0.313	0.610

Table 2: Maximum of all pair-wise euclidean distance between embeddings within an utterance (averaged across 300 random utterances). The small values in \mathbf{h} suggest similar embeddings.

MAE Target	<i>a</i> input	<i>a</i> output	<i>t</i> input	<i>t</i> output
random init	2.2	5.0	2.0	4.8
projected (\mathbf{h})	0.016	0.0014	0.008	0.0013
original (\mathbf{H})	4.3	3.0	4.0	2.7

Table 3: Retrieval accuracy ($N = 32$) and emotion recognition UA for different positional information types.

	Retrieval (%)		UA (%)		
	<i>a</i>	<i>t</i>	<i>a</i>	<i>t</i>	<i>a, t</i>
RoPE w/o contr	6.1	5.5	64.2	54.4	67.6
RoPE w/o <i>g</i>	88.0	91.6	63.6	54.5	67.2
Absolute PE	34.3	35.2	65.3	54.8	67.3
RoPE	42.5	44.8	65.1	55.0	67.9

ing that the "Separate", "Cascade" and "Late Fusion" all require multiple models for the different modality combinations, a common approach in production. Interestingly, the shared transformer with contr+MAE improved upon the baselines for both unimodal and multimodal cases, which highlights the usefulness of the reconstruction and modality alignment capability embedded within the shared transformers.

Third, the ablation in the last four rows suggests that the interplay between the contrastive and MAE losses is crucial. MAE alone in the shared transformer is not beneficial and led to a slight degradation, but combining the contrastive and MAE losses tends to further improve the performance. We conjecture that contrastive learning improved the alignment between modalities, which may allow masks from one modality to attend to those of another, and in turn further improving the multimodal knowledge.

4.4. Best Practices with MAE and Contrastive Learning

We found that the choices of target embeddings for MAE (Table 2) and positional embeddings for contrastive learning (Table 3) are crucial to ensure stabilized training.

Table 2 demonstrated that choosing the projected embeddings $\mathbf{h}_{i,j}^m$ as the target embeddings led to model collapse, whereas reconstructing the original embeddings $\mathbf{H}_{i,j}^m$ enabled stabilized training. This is because when reconstructing $\mathbf{h}_{i,j}^m$, the projection in Eq. (2) takes a *shortcut* for model collapse, so that the projected embeddings become nearly identical and can be "perfectly" reconstructed.

For the contrastive learning loss, the type of position information turned out to be integral, shown in Table 3. We found that absolute sinusoidal positional embeddings [17, 18] turn to have very large values that dominate output embeddings. This makes the contrastive objective a trivial task and difficult to optimize, as shown by retrieval accuracies in Table 3, defined as $\frac{1}{N} \sum_i 1\{\arg\max_i [\mathbf{z}_i^a \cdot (\mathbf{z}_i^t)^T] = i\}$ for *t* and similarly for *a*. In contrast, RoPE (which uses relative position information), does not have such issues. As shown in Table 3, not only is retrieval better with RoPE, but also the UA. Moreover, removing *g* led to overfitting to the contrastive objective and negative impact on UA, especially for the audio modality.

5. Conclusions

In this work, we tackle the problem of missing modalities in emotion recognition via a unified modality-agnostic shared transformer that is equipped with the capacity to do implicit reconstruction and modality alignment. Such model can support arbitrary combinations of input modalities. We show how the unified design of the model allows us to further improve its multimodal capabilities with carefully-designed additional objectives during training, and identify the major factors for their success, namely the choices of positional embeddings and target of the MAE reconstruction. The proposed system yielded state-of-the-art performance, compared with the literature and the implemented benchmarking baselines. Future work involves pre-training the shared transformers on larger datasets for enhanced capability for reconstruction and modality alignment.

6. References

- [1] G. Chochlakis, G. Mahajan, S. Baruah, K. Burghardt, K. Lerman, and S. Narayanan, "Leveraging label correlations in a multi-label setting: A case study in emotion," in *ICASSP*. IEEE, 2023, pp. 1–5.
- [2] S. Srinivasan, Z. Huang, and K. Kirchhoff, "Representation learn-

- ing through cross-modal conditional teacher-student training for speech emotion recognition,” in *ICASSP*. IEEE, 2022, pp. 6442–6446.
- [3] L. Goncalves, S.-G. Leem, W.-C. Lin, B. Sisman, and C. Busso, “Versatile audio-visual learning for handling single and multi modalities in emotion regression and classification tasks,” *arXiv preprint arXiv:2305.07216*, 2023.
 - [4] G. Chochlakakis, C. Lavania, P. Mathur, and K. Han, “Tackling missing modalities in audio-visual representation learning using masked autoencoders,” in *Interspeech*, 2024.
 - [5] O. Chang, O. Braga, H. Liao, D. Serdyuk, and O. Siohan, “On robustness to missing video for audiovisual speech recognition,” *TMLR*, 2022.
 - [6] J. Zhao, R. Li, and Q. Jin, “Missing modality imagination network for emotion recognition with uncertain missing modalities,” in *Proceedings of the 59th Annual Meeting of the ACL and the 11th International Joint Conference on NLP (Volume 1: Long Papers)*, Aug. 2021, pp. 2608–2618.
 - [7] Y. Wang, Y. Li, and Z. Cui, “Incomplete multimodality-diffused emotion recognition,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.
 - [8] M. Li, D. Yang, and L. Zhang, “Towards robust multimodal sentiment analysis under uncertain signal missing,” *IEEE Signal Processing Letters*, vol. 30, pp. 1497–1501, 2023.
 - [9] R. Huan, G. Zhong, P. Chen, and R. Liang, “Unimf: a unified multimodal framework for multimodal sentiment analysis in missing modalities and unaligned multimodal sequences,” *IEEE Transactions on Multimedia*, 2023.
 - [10] R. Liu, H. Zuo, Z. Lian, B. W. Schuller, and H. Li, “Contrastive learning based modality-invariant feature acquisition for robust multimodal emotion recognition with missing modalities,” *IEEE Transactions on Affective Computing*, 2024.
 - [11] H. Zuo, R. Liu, J. Zhao, G. Gao, and H. Li, “Exploiting modality-invariant feature for robust multimodal emotion recognition with missing modalities,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
 - [12] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, “Hubert: Self-supervised speech representation learning by masked prediction of hidden units,” *IEEE/ACM transactions on audio, speech, and language processing*, vol. 29, pp. 3451–3460, 2021.
 - [13] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized bert pretraining approach,” *arXiv preprint arXiv:1907.11692*, 2019.
 - [14] Z. Lian, L. Chen, L. Sun, B. Liu, and J. Tao, “Gcnet: Graph completion network for incomplete multimodal learning in conversation,” *IEEE Transactions on pattern analysis and machine intelligence*, vol. 45, no. 7, pp. 8419–8432, 2023.
 - [15] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual models from natural language supervision,” in *ICML*. PMLR, 2021, pp. 8748–8763.
 - [16] K. Avramidis, S. Stewart, and S. Narayanan, “On the role of visual context in enriching music representations,” in *ICASSP*, 2023, pp. 1–5.
 - [17] Y. Gong, A. Rouditchenko, A. H. Liu, D. Harwath, L. Karlinsky, H. Kuehne, and J. Glass, “Contrastive audio-visual masked autoencoder,” *arXiv preprint arXiv:2210.07839*, 2022.
 - [18] P. Wang, S. Wang, J. Lin, S. Bai, X. Zhou, J. Zhou, X. Wang, and C. Zhou, “One-peace: Exploring one general representation model toward unlimited modalities,” *arXiv preprint arXiv:2305.11172*, 2023.
 - [19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the NAACL: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.
 - [20] W. Kim, B. Son, and I. Kim, “Vilt: Vision-and-language transformer without convolution or region supervision,” in *International conference on machine learning*. PMLR, 2021, pp. 5583–5594.
 - [21] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.
 - [22] Y. Li, P. Bell, and C. Lai, “Speech emotion recognition with asr transcripts: A comprehensive study on word error rate and fusion techniques,” *arXiv preprint arXiv:2406.08353*, 2024.
 - [23] N. Das, S. Dingliwal, S. Ronanki, R. Paturi, Z. Huang, P. Mathur *et al.*, “Speechverse: A large-scale generalizable audio language model,” *arXiv preprint arXiv:2405.08295*, 2024.
 - [24] P. Mote, B. Sisman, and C. Busso, “Unsupervised domain adaptation for speech emotion recognition using k-nearest neighbors voice conversion,” in *Interspeech*, 2024.
 - [25] L. Goncalves, D. Robinson, E. Richerson, and C. Busso, “Bridging emotions across languages: Low rank adaptation for multilingual speech emotion recognition,” in *Interspeech*, 2024.
 - [26] A. R. Naini, M. A. Kohler, E. Richerson, D. Robinson, and C. Busso, “Generalization of self-supervised learning-based representations for cross-domain speech emotion recognition,” in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 12 031–12 035.
 - [27] R. Lotfian and C. Busso, “Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings,” *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, 2017.
 - [28] J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, and Y. Liu, “Roformer: Enhanced transformer with rotary position embedding,” *Neurocomputing*, vol. 568, p. 127063, 2024.
 - [29] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” in *ICML*. PMLR, 2023, pp. 28 492–28 518.
 - [30] S. Chen *et al.*, “Wavlm: Large-scale self-supervised pre-training for full stack speech processing,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.