# MEMOBERT: PRE-TRAINING MODEL WITH PROMPT-BASED LEARNING FOR MULTIMODAL EMOTION RECOGNITION

*Jinming Zhao*[1,2], *Ruichen Li*[1], *Qin Jin*[1]*, *Xinchao Wang*[2], *Haizhou Li*[2,3]

[1] School of Information, Renmin University of China
[2] Electrical and Computer Engineering, National University of Singapore
[3] The Chinese University of Hong Kong, Shenzhen, China

**ABSTRACT**

Multimodal emotion recognition study is hindered by the lack of labelled corpora in terms of scale and diversity, due to the high annotation cost and label ambiguity. In this paper, we propose a multimodal pre-training model **MEmoBERT** for multimodal emotion recognition, which learns multimodal joint representations through self-supervised learning from a self-collected large-scale unlabeled video data that come in sheer volume. Furthermore, unlike the conventional "pre-train, finetune" paradigm, we propose a prompt-based method that reformulates the downstream emotion classification task as a masked text prediction one, bringing the downstream task closer to the pre-training. Extensive experiments on two benchmark datasets, IEMOCAP and MSP-IMPROV, show that our proposed MEmoBERT significantly enhances emotion recognition performance.

***Index Terms***— Emotion Recognition, Multimodal, Pre-training, Prompt

# 1 Introduction

Automatic Multimodal Emotion Recognition aims to interpret human emotions through multiple modalities and serves as an enabling technology for many applications [1, 2]. Previous works have explored multimodal fusion strategies and achieved superior performance on multimodal emotion recognition tasks. For example, MFN [3] and MARN [4] are attention- and memory-based approaches applicable to sequences aligned at the word level. MulT [5] is a transformer-based framework handling non-aligned multimodal sequences and long-range dependencies. However, these methods are limited by the scarcity of the labeled data.

In recent years, various pre-trained models via self-supervised learning on large-scale unlabeled data have achieved promising results. The pre-trained language models, such as BERT [6], ELMo [7] and GPT [8], have attracted much attention and are widely adopted. Inspired by the success of pre-training in text modality, many image&language cross-modality pre-trained models are proposed [9, 10, 11] and have achieved new state-of-the-art performances on various tasks, such as Image-Text Retrieval, Visual Question Answering, etc. Such models, including UNITER [11] and LXMERT [12], typically follow a single- or dual-stream multi-layer Transformer architecture and optimize through several pre-training tasks to learn multimodal joint representations, such as Masked Language Modeling (MLM), Masked Region Modeling (MRM), Image-Text Matching (ITM). VideoBERT [13], ActBERT [14] and HERO [15] extend image&language representation learning to the video&language representation learning for the video-related tasks, such as Video-Text Retrieval, Video Action Classification, etc. Furthermore, VATT [16] considers three modalities (text, visual and audio) with a modality-agnostic Transformer and uses multimodal contrastive learning to learn multimodal representations.

Prompt-based learning [17], on the other hand, has achieved great success and has become a new learning paradigm in NLP. Compared to the "pre-train, finetune" paradigm, which adapts pre-trained models to downstream tasks via objective engineering, the "pre-train, prompt and predict" paradigm reformulates the downstream tasks to resemble the masked language modeling task optimized in the pre-training phase with the help of a prompt. This paradigm brings the downstream tasks closer to the pre-training tasks, which can retain more knowledge learned during pre-training for downstream tasks. It outperforms the "pre-train, finetune" paradigm, especially under low-resource conditions in many NLP tasks [18, 19, 20, 21].

Motivated by the above studies, we propose a multimodal transformer-based pre-training model, MEmoBERT, to learn multimodal joint representations for emotion recognition. It is trained through self-supervised learning based on a large-scale unlabeled video dataset containing more than 350 movies. We design four efficient self-supervised pre-training tasks to learn multimodal joint representations, including Whole Word Masked Language Modeling (WWMLM), Span Masked Visual Frame Modeling (SpanMVFM), Span Masked Visual Frame Classification with KL-divergence (SpanMVFC-KL), Span Masked Acoustic Frame Modeling (SpanMAFM). Furthermore, we explore a prompt-based

*Corresponding author.

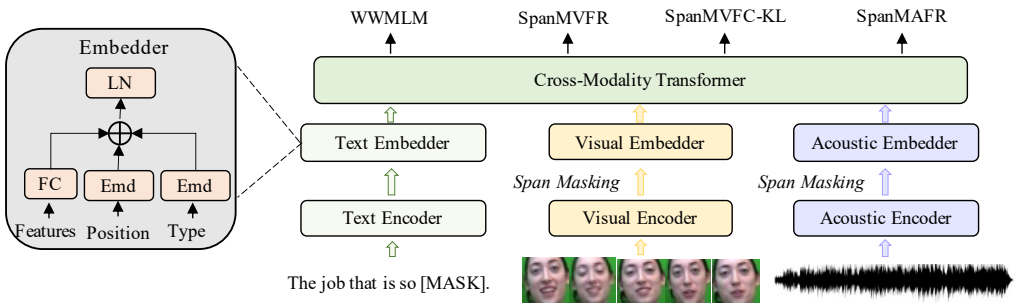



**Fig. 1**. Overview of the proposed MEmoBERT model, consisting of three modality-specific Encoders, three modality-specific Embedders and a multi-layer Cross Modality Transformer, learned through four pre-training tasks. In the Embedder, "FC" and "Emd" refer to a fully-connected layer and a embedding layer respectively, and "Type" refers to different modalities (e.g. Type 0 for text modality). Different modality Embedders have the same structure but independent parameters.

learning method to efficiently adapt the pre-trained model to downstream tasks. To the best of our knowledge, MEmoBERT is the first multimodal pre-training model for multimodal emotion recognition, and is also the first to adopt prompt-based learning for this task. We carry out experiments on two benchmark datasets, IEMOCAP and MSP-IMPROV, to evaluate our pre-trained MEmoBERT under full- and partial-training data conditions. The results show that our MEmoBERT yields significant improvement and the prompt-based method further improves the performance.

The main contributions of this work include, 1) We propose a multimodal transformer-based pre-trained model, MEmoBERT, with self-supervised learning on a large-scale unlabeled video dataset for multimodel emotion recognition. 2) We propose a prompt-based learning method that better adapts the pre-trained MEmoBERT to downstream tasks. 3) MEmoBERT achieves a new state-of-the-art performance on both multimodal emotion recognition benchmark datasets.

# 2 Method

Fig. 1 illustrates the overall model framework of our proposed MEmoBERT and its learning process during pre-training. MEmoBERT consists of three independent modality Encoders to generate modality-specific token/frame-level raw features for the text, visual, and acoustic modalities, and three Embedders to generate embeddings based on corresponding raw features, type and positions of each modality respectively. Specifically, the embedding layer in BERT [6] is adopted as the Text Encoder. The Visual Encoder is a pre-trained facial expression model that generates the facial expression features/distributions based on the speaker faces. The Acoustic Encoder is a pre-trained speech model that generates the acoustic features based on the audio waveform. The final embedding for each modality is obtained via its modality Embedder which sums up the raw features, position embeddings and type embeddings, and then gets normalized via Layer Norm. Please note that the parameters of Acoustic Encoder and Visual Encoder are fixed during pre-training. A cross-modality transformer in MEmoBERT then learns cross-modality contextualized representation based on the embeddings from different modalities.

We design four efficient pre-training tasks to optimize MEmoBERT in the pre-training phase to learn emotional multimodal joint representations. Once the model is well pre-trained, we adopt the prompt-based or finetune-based method to adapt it to downstream tasks.

## 2.1 Cross Modality Transformer

The cross-modality transformer adopts the most established Transformer architecture [6] and extends it to three modalities (text, visual and audio) for multimodal pre-training. We follow the modality-agnostic strategy [11, 16], that is, a single backbone Transformer is applied to any of the modalities. During pre-training, the modality-specific embeddings are fed into the multi-layer Transformer to learn high-level cross-modality contextualized representations across different modalities.

## 2.2 Pre-training Tasks

We design four pre-training tasks including text, visual and audio modality-related tasks to enhance the cross-modality interaction and to learn multimodal joint emotional representations. In all following pre-training tasks, we adopt the conditional masking strategy which only masks one modality and keeps other modalities intact in corresponding tasks. It can learn the better latent alignment across three modalities and enables the model to learn better multimodal joint representations [11]. For example, as shown in Fig. 1, in the case where the word "cool" is masked (WWMLM), our model should be able to infer the masked word based on the surrounding text, facial frames and acoustic signals. While in the case where the several smiley faces are masked (SpanMVFR/SpanMVFC-KL), our model should be able to infer the facial expressions features/emotion distribution of the masked facial frames based on the surrounding facial expressions, the word "cool", and the voice tone.

**Whole Word Masked Language Modeling (WWMLM)** learns to predict the masked whole words conditioned on the visual and the acoustic modalities. The whole word mask-

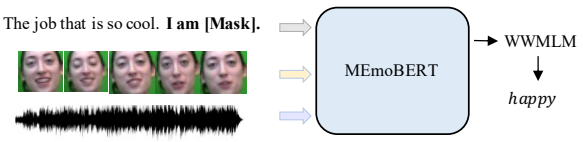



**Fig. 2**. "prompt, predict" paradigm based on the pre-trained MEmoBERT for downstream tasks.

ing strategy that masks whole words rather than WordPiece tokens can better capture the accurate semantics [22]. For example, masking partial WordPiece tokens of a word may lead to completely opposite emotional semantics of the whole word, especially for words with the prefixes (e.g. "un-", "im-", "op-") and the suffixes (e.g. "-less" ).

**Span Masked Acoustic Frame Regression (SpanMAFR)** learns to reconstruct the acoustic features extracted from the acoustic encoder of the masked audio frames conditioned on the text and visual modalities. Furthermore, we adopt the span masking strategy [23, 24] which masks consecutive frames. It aims to capture global emotional expression and avoid the model exploiting the local smoothness of acoustic frames. We apply L2 regression as the objective function to minimize the reconstruction error between the predicted and ground-truth frames.

**Span Masked Visual Frame Regression (SpanMVFR)** learns to reconstruct the facial expression features extracted from the visual encoder of the masked visual frames conditioned on the text and acoustic modalities. Due to the similarity of consecutive visual frames, similar to that in the acoustic modality, we also adopt the span masking strategy for the visual modality. We also apply L2 regression as the objective function to minimize the reconstruction error.

**Span Masked Visual Frame Classification with KL-divergence (SpanMVFC-KL)** learns to predict the distribution of the emotion categories (such as happy, sad, anger) for each masked visual frame conditioned on the text and acoustic modalities. We feed the Transformer output of the masked frame into a fully connected layer to predict the emotion distribution of $K$ facial expression classes. Finally, we use the KL-divergence objective function to optimize the predicted emotion distributions with respect to the ground-truth emotion distributions produced by the visual encoder.

## 2.3 Prompt-based Emotion Classification

Fig. 2 illustrates the "prompt, predict" paradigm. We extend the prompt-based learning in NLP to multimodal scenarios by simply add a prompt following the text modality. Given a prompt-based multimodal input "[X] I am [MASK]. [V] [A]" where the $[X], [V], [A]$ are text, visual and acoustic inputs of the video respectively, the classification problem is thus reformulated to predict the "[MASK]" as an emotion category word (such as happy, sad, anger) with the help of a textual prompt "I am [MASK].". It is similar to the masked language modeling task in the pre-training phase.

| dataset | Happy | Anger | Sadness | Neutral | Total |
|---|---|---|---|---|---|
| IECMOAP | 1636 | 1103 | 1084 | 1708 | 5531 |
| MSP-IMPROV | 999 | 460 | 627 | 1733 | 3819 |

**Table 1**. Statistics of the benchmark datasets.

# 3 Experiments

## 3.1 Pre-training Dataset

Providing a large-scale unlabeled dataset is one of the core challenges for pre-training methods. We collect 351 movies and TV series in the categories of family and romance, which have rich and natural emotional expressions. We extract the speakers' faces, audio, and subtitles of each utterance and filter out the empty utterances. In the end, we build a dataset containing about 180k utterances with three modalities.

## 3.2 Benchmark Datasets

We evaluate our proposed MEmoBERT model on two multimodal emotion recognition benchmark datasets: IEMOCAP [25] and MSP-IMPROV [26]. We follow the emotional label processing in [27] for the four-class emotion recognition setup. The statistics of the datasets are shown in Table 1.

## 3.3 Implementation Details

### 3.3.1 Modality Encoders

**Acoustic Encoder:** We adopt a SOTA pre-trained speech model, Wav2Vec2.0 [28], as the acoustic encoder to extract the frame-level acoustic features. We down-sample the frame-level features by average pooling every 3 frames.

**Visual Encoder:** We adopt a pre-trained facial expression recognition model, DenseFace, as the visual encoder, which has a DenseNet structure [29] and is trained on a facial expression corpus, FER+ [30]. Then, the encoder is used to extract the frame-level facial expression features and emotion category distributions based on the speaker's faces detected based on the consistency of voice activation and mouth movement. We down-sample the frame-level features every 3 frames.

### 3.3.2 Experiment Setups

During MEmoBERT pre-training, we first initialize its weights from a text pre-trained BERT checkpoint[1]. Specifically, MEmoBERT uses the same backbone architecture as BERT. For text modality, we follow the masking strategy used in BERT [6]. For the visual and acoustic modality, we follow the span masking strategy [23] with the masking ratio of 15% and the consecutive masking number of 3. We use AdamW optimizer with initial learning rate of 5e-5 over maximum 40K steps. The batch size is 640.

For downstream tasks, we use the 10-fold and 12-fold speaker-independent cross-validation to evaluate the models on IEMOCAP and MSP-IMPROV respectively. In each fold, we use one speaker for testing and the remaining speakers for training. We use the weighted accuracy (WA) and unweighted average recall (UAR) as the evaluation metrics. We

[1]https://huggingface.co/bert-base-uncased

| | IEMOCAP | | MSP-IMPROV | |
|---|---|---|---|---|
| | WA | UAR | WA | UAR |
| cLSTM-MMA [31] | 73.94% | – | – | – |
| SSMM [32] | 75.60% | 74.50% | – | – |
| MMIN [27] | – | 78.12% | – | 68.55% |
| Direct | 74.64% | 75.76% | 67.17% | 65.57% |
| BERT+Finetune | 77.98% | 78.98% | 70.08% | 69.67% |
| Pretrain+Finetune | 79.63% | 80.61% (+1.6) | 71.77% | 71.35% (+1.7) |
| Pretrain+Prompt | 80.01% | 81.09% (+2.1) | 72.36% | 72.22% (+2.5) |

**Table 2**. Performance comparison on IEMOCAP and MSP-IMPROV. The numbers in blue refer to the improvements compared to "BERT+Fintune".

run three times for each experiment and report the average performance. We set the initial learning rate as 5e-5 and 3e-5 for experiments on full- and partial-training data respectively over maximum 15 epochs. The batch size is 32.

In order to verify the effectiveness of our model framework and the prompt-based learning method, we specifically define four experiment settings: 1) "Direct" denotes that we directly train the MEmoBERT followed by a classifier for downstream tasks from scratch. 2) "BERT+Finetune" denotes that we finetune the MEmoBERT followed by a classifier for downstream tasks, in which the MEmoBERT is initialized by the pre-trained BERT. 3) "Pretrain+Finetune" denotes that we finetune the pre-trained MEmoBERT followed by a classifier for downstream tasks. 4) "Pretrain+Prompt" denotes that we adopt the prompt-based learning method based on the pre-trained MEmoBERT without introducing any additional parameters for downstream tasks.

### 3.4 Experiments Results

Table 2 presents the multimodal emotion recognition results on the two benchmark datasets, IEMOCAP and MSP-IMPROV. Compared to other state-of-the-art models without pre-training in the first block, "BERT+Finetune" achieves superior performance on both datasets, which demonstrates that the pre-trained language modal BERT can benefit the multimodal emotion recognition. "Pretrain+Finetune" based on our pre-trained MEmoBERT achieves significant improvement compared to "BERT+Finetune". Furthermore, "Pretrain+Prompt" with prompt-based learning over our pre-trained MEmoBERT can bring additional improvement.

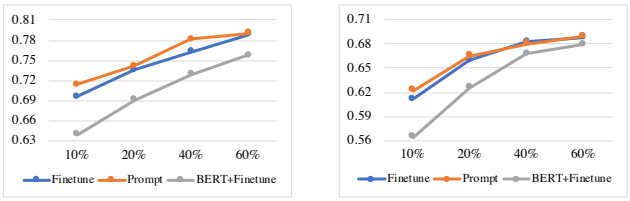
### 3.5 Ablation Study

**Ablation of the pre-training tasks** We first investigate the impact of different pre-training tasks on the performance of MEmoBERT. As shown in Table 3, all pre-training tasks and strategies, including the span masking and whole word masking strategies, and the visual and acoustic related pre-training tasks, are beneficial for pre-training MEmoBERT.

**Ablation of different amounts of training data.** In order to validate the robustness of the pre-trained MEmoBERT and prompt-based method under low-resource conditions, we conduct experiments using different amounts of training data in the downstream tasks. As shown in Fig. 3, applying "Fine-tune" and "Prompt" on the pre-trained MEmoBERT both significantly outperforms the "BERT+Finetune" setting under all low-resource conditions, and the less training data, the more obvious the improvement brought by the pre-trained MEmoBERT. The prompt-based method outperforms the finetune-based method on IEMOCAP and MSP under almost all conditions. It indicates that the prompt-based method can efficiently adapt the pre-trained MEmoBERT to downstream tasks, especially under low-resource conditions.

| | IEMOCAP | | MSP-IMPROV | |
|---|---|---|---|---|
| | WA | UAR | WA | UAR |
| Pretrain+Prompt | 80.01% | 81.09% | 72.36% | 72.22% |
| w/o visual tasks | 79.73% | 80.89% | 71.04% | 70.88% |
| w/o acoustic task | 79.48% | 80.82% | 71.78% | 71.52% |
| w/o span&whole_word | 79.46% | 80.70% | 70.73% | 71.01% |

**Table 3**. Ablation study of the pre-training tasks. "visual tasks" refers to "SpanMVFR" and "SpanMVFC-KL". "acoustic task" refers to "SpanMAFR". "span&whole_word" denotes removing the whole word masking strategy in WWMLM and removing span masking strategy in "SpanMVFR", "SpanMVFC-KL" and "SpanMAFR" and using the simple random token/frame masking strategies.




(a) IEMOCAP (b) MSP-IMPROV

**Fig. 3**. Performance (UAR) comparison with different amount of training data. "10%" means only 10% of the training data are used for training.

## 4 Conclusion

In this paper, we propose a novel multimodal transformer-based pre-trained model, MEmoBERT, with self-supervised learning on a self-collected large-scale unlabeled video dataset for multimodal emotion recognition. We further investigate a prompt-based learning method that can better adapt the pre-trained MEmoBERT to downstream tasks. Extensive experiments on two public datasets demonstrate the effectiveness and robustness of our proposed methods.

## 5 Acknowledgments

This work was partially supported by the National Key RD Program of China (No.2020AAA0108600), the National Natural Science Foundation of China (No. 62072462), Large-Scale Pre-Training Program 468 of Beijing Academy of Artificial Intelligence, A*STAR RIE2020 Advanced Manufacturing and Engineering Domain (AME) Programmatic Grant (No. A1687b0033) and China Scholarship Council.

# 6 References


[1] Nickolaos Fragopanagos and John G Taylor, "Emotion recognition in human–computer interaction," *Neural Networks*, vol. 18, no. 4, pp. 389–405, 2005.

[2] Morena Danieli, Giuseppe Riccardi, and Firoj Alam, "Emotion unfolding and affective scenes: A case study in spoken conversations," in *Proceedings of the International Workshop on Emotion Representations and Modelling for Companion Technologies*, 2015, pp. 5–11.

[3] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, and et al., "Memory fusion network for multi-view sequential learning," in *AAAI*, 2018, vol. 32.

[4] Amir Zadeh, Paul Pu Liang, Soujanya Poria, and et al., "Multi-attention recurrent network for human communication comprehension," in *AAAI*, 2018.

[5] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, and et al., "Multimodal transformer for unaligned multimodal language sequences," in *ACL*. NIH Public Access, 2019, vol. 2019, p. 6558.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and et al., "Bert: Pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT (1)*, 2019.

[7] Matthew E Peters, Mark Neumann, Mohit Iyyer, and et al., "Deep contextualized word representations," in *Proceedings of NAACL-HLT*, 2018, pp. 2227–2237.

[8] Alec Radford, Jeffrey Wu, Rewon Child, and et al., "Language models are unsupervised multitask learners," *OpenAI Technical Report*.

[9] Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and et al., "Multimodal pretraining unmasked: A meta-analysis and a unified framework of vision-and-language berts," *ACL*, vol. 9, pp. 978–994, 2021.

[10] Weijie Su, Xizhou Zhu, Yue Cao, and et al., "Vl-bert: Pre-training of generic visual-linguistic representations," *arXiv preprint arXiv:1908.08530*, 2019.

[11] Yen-Chun Chen, Linjie Li, Licheng Yu, and et al., "Uniter: Universal image-text representation learning," 2020.

[12] Hao Tan and Mohit Bansal, "Lxmert: Learning cross-modality encoder representations from transformers," *arXiv preprint arXiv:1908.07490*, 2019.

[13] Chen Sun, Austin Myers, Carl Vondrick, and et al., "Videobert: A joint model for video and language representation learning," in *ICCV*, 2019, pp. 7464–7473.

[14] Linchao Zhu and Yi Yang, "Actbert: Learning global-local video-text representations," in *CVPR*, 2020, pp. 8746–8755.

[15] Linjie Li, Yen-Chun Chen, Yu Cheng, and et al., "Hero: Hierarchical encoder for video+ language omni-representation pre-training," in *EMNLP*, 2020, pp. 2046–2065.

[16] Hassan Akbari, Linagzhe Yuan, Rui Qian, and et al., "Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text," *arXiv preprint arXiv:2104.11178*, 2021.

[17] Pengfei Liu, Weizhe Yuan, Jinlan Fu, and et al., "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *arXiv preprint arXiv:2107.13586*, 2021.

[18] Fabio Petroni, Tim Rocktäschel, Patrick Lewis, and et al., "Language models as knowledge bases?," *arXiv preprint arXiv:1909.01066*, 2019.

[19] Ruiqi Zhong, Kristy Lee, Zheng Zhang, and et al., "Adapting language models for zero-shot learning by meta-tuning on dataset and prompt collections," 2021.

[20] Timo Schick and Hinrich Schütze, "Few-shot text generation with pattern-exploiting training," *arXiv preprint arXiv:2012.11926*, 2020.

[21] Timo Schick and Hinrich Schütze, "Exploiting cloze-questions for few-shot text classification and natural language inference," in *ACL*, 2021, pp. 255–269.

[22] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu, "Pre-training with whole word masking for chinese bert," *arXiv preprint arXiv:1906.08101*, 2019.

[23] Andy T Liu, Shu-wen Yang, Po-Han Chi, and et al., "Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders," in *ICASSP*. IEEE, 2020, pp. 6419–6423.

[24] Dongwei Jiang, Wubo Li, Ruixiong Zhang, and et al., "A further study of unsupervised pretraining for transformer based speech recognition," in *ICASSP*. IEEE, 2021, pp. 6538–6542.

[25] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, and et al., "Iemocap: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[26] Carlos Busso, Srinivas Parthasarathy, Alec Burmania, and et al., "Msp-improv: An acted corpus of dyadic interactions to study emotion perception," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, 2016.

[27] Jinming Zhao, Ruichen Li, and Qin Jin, "Missing modality imagination network for emotion recognition with uncertain missing modalities," in *ACL*, 2021, pp. 2608–2618.

[28] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and et al., "wav2vec 2.0: A framework for self-supervised learning of speech representations," *arXiv preprint arXiv:2006.11477*, 2020.

[29] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and et al., "Densely connected convolutional networks," in *CVPR*, 2017, pp. 4700–4708.

[30] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and et al., "Training deep networks for facial expression recognition with crowd-sourced label distribution," in *ICMI*, 2016, pp. 279–283.

[31] Zexu Pan, Zhaojie Luo, Jichen Yang, and et al., "Multi-modal attention for speech emotion recognition," *Interspeech*, pp. 364–368, 2020.

[32] Jingjun Liang, Ruichen Li, and Qin Jin, "Semi-supervised multi-modal emotion recognition with cross-modal distribution matching," in *ACM Multimedia*, 2020, pp. 2852–2861.