EmotionNAS: Two-stream Neural Architecture Search for Speech Emotion Recognition

Haiyang Sun[1,2*], *Zheng Lian*[2*], *Bin Liu*[2†], *Ying Li*[2], *Licai Sun*[1,2], *Cong Cai*[1,2], *Jianhua Tao*[3†], *Meng Wang*[4], *Yuan Cheng*[4]

[1]School of Artificial Intelligence, University of Chinese Academy of Sciences
[2]State Key Laboratory of Multimodal Artificial Intelligence Systems
[3]Department of Automation, Tsinghua University, [4]Ant Group

{sunhaiyang2021, lianzheng2016}@ia.ac.cn

Abstract

Speech emotion recognition (SER) is an important research topic in human-computer interaction. Existing works mainly rely on human expertise to design models. Despite their success, different datasets often require distinct structures and hyperparameters. Searching for an optimal model for each dataset is time-consuming and labor-intensive. To address this problem, we propose a two-stream neural architecture search (NAS) based framework, called "EmotionNAS". Specifically, we take two-stream features (i.e., handcrafted and deep features) as the inputs, followed by NAS to search for the optimal structure for each stream. Furthermore, we incorporate complementary information in different streams through an efficient information supplement module. Experimental results demonstrate that our method outperforms existing manually-designed and NAS-based models, setting the new state-of-the-art record.

Index Terms: speech emotion recognition, neural architecture search, two-stream model, information supplement module.

1. Introduction

Speech emotion recognition (SER) has received increasing attention due to its contribution to human-computer interactions [1, 2, 3]. SER aims to understand how humans express their emotions and then classify each utterance into its emotional state [4, 5]. Existing works are mainly manually-designed models [6, 7, 8, 9]. Despite their success, these works rely on historical experience to design model structures, which is often time-consuming and labor-intensive [10, 11]. Therefore, how to design networks more intelligently has been brought into focus. To this end, we explore neural architecture search (NAS) for SER [12, 13]. By setting the search space, the search strategy, and the evaluation metric, we can optimize the model architecture automatically with little human intervention.

Previously, search methods based on reinforcement learning [14] or evolutionary algorithms [15] avoided the exhaustive search and provided deep learning-inspired design principles. However, these methods need the retraining process every time a new substructure is sampled, which consumes numerous computational resources. To speed up the search process, researchers propose differentiable architecture search (DARTS) [10, 16, 17, 18], which relaxes the search space to be continuous by applying a softmax operation. Additionally, some efforts [19, 20] build a model library by pre-training multiple structures. These methods successfully alleviate the time-consuming process in previous works. In this paper, we rely on more efficient search methods to design SER models.

*Equal contribution †Corresponding Author

In addition to model design, another challenge in SER is how to represent audios [21]. Among existing handcrafted features, spectrograms are widely utilized in SER [22, 23, ?, 24]. However, the spectrogram loses some phase information during calculation. Recently, the deep feature, wav2vec [25], has demonstrated its effectiveness in speech representation learning [26]. It is a self-supervised framework that can learn powerful acoustic representations with the help of large amounts of unlabeled data. However, wav2vec may lose some emotion-related information due to different training objectives. To obtain a more comprehensive speech representation, we integrate handcrafted and deep features via a two-stream framework.

Meanwhile, how to fuse multiple features also affects the classification performance. Unsuitable fusion approaches cannot effectively utilize the complementary information in different features, which may lead to performance degradation compared to the best-performing feature (denoted as the dominant feature in this paper). Therefore, we further design an effective fusion method, called "information supply module (ISM)", to ensure that the performance of the dominant feature is not degraded and can be further improved.

In summary, to address the low efficiency of the manually-designed approach and inefficient fusion of different features, we propose a novel two-stream framework called "EmotionNAS". Figure 1 shows the overall structure of our proposed method. Experimental results show that EmotionNAS outperforms existing manually-designed and NAS-based models. The contributions of this paper can be summarized as follows:

- We propose a novel framework called "EmotionNAS". It incorporates complementary information in handcrafted and deep features, followed by NAS to efficiently search through numerous possible networks to find the optimal structure.
- We further design ISM to effectively fuse complementary information in different features.
- Experimental results on IEMOCAP show that our method successfully outperforms existing manually-designed and NAS-based models, setting the new state-of-the-art record.

2. Methodology

EmotionNAS is a two-stream framework that combines handcrafted and deep features for SER. To search for the optimal structure, we explore NAS in model design. Subsequently, we employ ISM to fuse two-stream features effectively.

2.1. Architecture Search for Spectrogram

To better utilize the time-frequency two-dimensional information in the spectrogram, our method automatically optimizes the CNN-based structure through an efficient architecture search al-
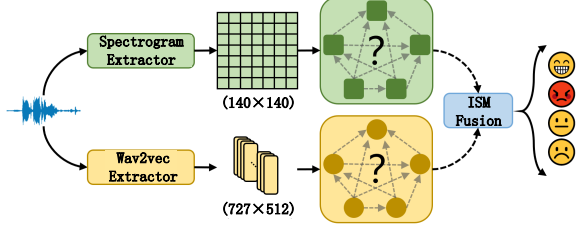
Figure 1: *The overall structure of EmotionNAS. It is a two-stream architecture that takes spectrogram and wav2vec as the inputs, followed by NAS to design the model automatically. We further fuse the outputs of two branches through ISM to achieve better performance.*
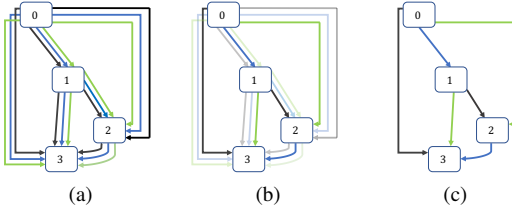

Figure 2: *DARTS architecture search in a cell: (a) the initial state of a cell with an ordered sequence of $N = 4$ nodes and $|\mathbb{O}| = 3$ operations. Different-colored arrows represent distinct operations; (b) during training, some operations' weights are reduced, resulting in lighter-colored lines; (c) only the operations with the highest weight are retained.*

gorithm DARTS [16]. It divides the whole network into normal and reduction cells. All operations adjacent to the reduction cells have a stride of 2, while others have a stride of 1.

Figure 2 shows the overall search process in a cell. Suppose each cell consists of an ordered sequence of N nodes. Let $x^{(i)}$ denotes the latent feature of the i-th node. Assume $\mathbb{O}$ is a set of candidate operations, and $|\mathbb{O}|$ represents the number of operations. To make the search space continuous, DARTS incorporates all possible operations through weights. We take the operations from $x^{(i)}$ to $x^{(j)}$ as an example:

$$\alpha_o^{(i,j)} = \text{softmax}\left(\theta_o^{(i,j)}\right), \qquad (1)$$

$$x^{(j)} = \sum_{o \in \mathbb{O}} \alpha_o^{(i,j)} o(x^{(i)}), \qquad (2)$$

where $\theta_o^{(i,j)}$ and $\alpha_o^{(i,j)}$ represent the raw and normalized weights of the operation $o \in \mathbb{O}$ from $x^{(i)}$ to $x^{(j)}$, respectively. The task of architecture search reduces to learning a set of continuous weights $\{\alpha_o^{(i,j)}\}$. At the end of the search, we only keep the operation with the highest weight.

2.2. Architecture Search for Wav2vec

Since wav2vec contains temporal information, we search for the optimal RNN-based structure for this branch [20]. The search space includes all regular operations in RNN, such as linear mapping, blending, activation function, and element-wise operation. The optimized cell structure is shared across different timesteps. For timestep t, the initial nodes consist of the input vector x_t and two hidden states h_{t-1}^1 and h_{t-1}^2. We aim to optimize the cell architecture and generate new hidden states for

the next timestep, i.e., h_t^1 and h_t^2 (see Section 4.6 for details).

At the end of the search, we feed wav2vec into the optimized recurrent neural architecture to generate frame-level representations. Since different frames play different parts in SER, we further exploit an attention mechanism to prioritize important frames and fuse them [27].

2.3. Information Supplement Module

After finding the optimal structure for each stream, we propose an ISM to fuse complementary information from two streams. To ensure that the performance of the dominant feature is not degraded after fusion, we abandon to compress this feature. Instead, we select complementary information from other inputs and fuse them with the dominant feature.

Specifically, suppose the outputs of two branches are denoted as $X_1 \in \mathbb{R}^{d_1}$ and $X_2 \in \mathbb{R}^{d_2}$, respectively. Here, d_1 and d_2 are the feature dimensions of X_1 and X_2. Suppose X_2 is the dominant feature. The calculation formula of our ISM can be summarized as follows:

$$H_1 = \text{Proj}\left(X_1\right), \qquad (3)$$

$$A_1 = \tanh\left(\text{Proj}\left(X_2\right) \odot H_1\right), \qquad (4)$$

$$F = X_2 + A_1 \odot H_1, \qquad (5)$$

where $\text{Proj}(\cdot)$ and $\odot$ represent a linear transformation function and an element-wise multiplication, respectively. Here, $H_1 \in \mathbb{R}^{d_2}$ and $A_1 \in \mathbb{R}^{d_2}$. In the end, we leverage the fused feature $F \in \mathbb{R}^{d_2}$ for emotion recognition.

3. Experimental Database and Setup

We first describe the benchmark dataset in our experiments. Following that, we illustrate the feature extraction process, implementation details, and various current advanced baselines.

3.1. Database

IEMOCAP [28] is a benchmark dataset for SER. It consists of five sessions, each with two speakers. For a fair comparison, we adopt five-fold cross-validation with the leave-one-session-out strategy [9]. Eight speakers from four sessions are used as the training set. One speaker in the remaining session is used as the validation set and the other as the test set. We evaluate four emotions (i.e., *neutral*, *angry*, *happy* and *sad*) on improvised data, in line with previous works [9]. Due to the imbalanced class distribution, we use unweighted accuracy (UA) as the primary metric and also report weighted accuracy (WA).

3.2. Feature Extraction

In EmotionNAS, we take spectrogram and wav2vec as inputs. The extraction processes are described as follows:

Spectrogram: We utilize the librosa toolkit [29] to extract spectrograms from audio. We unify the duration of audio into 8 seconds by zero padding and truncation. Then, a spectrogram is extracted with 25ms Hamming windows and 14ms overlap. Finally, we downsample the spectrogram to (140×140) using an average pooling operation.

Wav2vec: We use the pretrained *wav2vec-large* [25] as the acoustic feature extractor. We unify the wav2vec feature to the maximum length by zero padding, followed by average pooling to downsample it to (727×512).

Table 1: *Performance of different approaches. Bold font represents the best performance.*

Method	UA(%)	WA(%)
CNN-GRU [6]	64.2	71.5
CTC-RNN [7]	65.7	64.2
SeqCap [8]	59.7	72.7
PCNSE-SADRN-CTC [9]	66.3	**73.1**
UniformNAS [32]	56.9	70.5
EmotionNAS (Ours)	**69.1**	72.1

Table 2: *Performance comparison between NAS-based methods and ResNet. ∗ represents our fair comparison by adjusting the model structure with the similar number of parameters.*

Spectrogram Branch	Params	UA(%)	WA(%)
ResNet18	11.18M	**59.4**	60.7
ResNet*	1.23M	55.2	58.4
NAS-C4L3	0.07M	52.1	54.4
NAS-C6L3	0.13M	57.3	**63.2**
NAS-C8L4	0.35M	58.6	62.2

3.3. Implementation Details

Our method consists of three key modules: 1) NAS for spectrogram; 2) NAS for wav2vec; 3) ISM for fusion. Since wav2vec can achieve better performance in SER, we treat this feature as the dominant feature. For the search process, we set hyperparameters based on validation performance.

NAS for spectrogram: In each cell, we set the number of nodes $N = 4$ and the number of operations $|\mathbb{O}| = 8$. The search space contains regular operations in CNNs, such as 3×3 max pooling, 3×3 average pooling, skip connection, 3×3 separable convolution, 5×5 separable convolution, 3×3 dilated convolution, 5×5 dilated convolution, and no connection. At the same time, we set the initial number of channels $C = 6$ and the number of layers $L = 3$.

NAS for wav2vec: The search space contains regular operations in RNNs, such as linear function, blending, element-wise product, element-wise sum, tanh function, sigmoid function, and Leaky ReLU function. Previous work [20] provided various RNN-based cell structures searched on the Penn Tree Bank dataset [30]. We choose the structure with the lowest validation loss in the emotion dataset. Meanwhile, we set the number of layers $L = 1$ and the hidden feature dimension $H = 256$.

3.4. Baselines

To verify the effectiveness of our method, we compare the performance of EmotionNAS with various baselines: **CNN-GRU** [6] combines CNN and GRU to recognize emotional states. **CTC-RNN** [7] exploits connectionist temporal classification [31] to annotate labels for audio segments automatically. **SeqCap** [8] uses the capsule network and GRUs to capture spatial-temporal information. **PCNSE-SADRN-CTC** [9] models long-range dependencies by combining parallel convolutional layers, squeeze-and-excitation networks, and self-attention dilated residual networks. **UniformNAS** [32] employs NAS to search the optimal structure, and uses a uniform path dropout strategy to encourage all structures to be searched equally.

Table 3: *Performance comparison between NAS-based method with RNN and LSTM.*

Wav2vec Branch	Params	UA(%)	WA(%)
RNN-H512L2	1.05M	60.2	64.1
RNN-H512L1	0.53M	61.0	64.4
LSTM-H512L2	4.21M	64.6	65.7
LSTM-H512L1	2.10M	64.6	67.7
NAS-H512L1	2.37M	**67.2**	68.4
NAS-H256L1	0.73M	66.2	**70.3**

Table 4: *Importance of each branch.*

Method	UA(%)	WA(%)
spectrogram branch	57.3	63.2
wav2vec branch	66.2	70.3
EmotionNAS (Ours)	**69.1**	**72.1**

4. Results and Discussion

In this section, we first conduct comparative experiments with currently advanced systems to verify the effectiveness of our method. Then, we systemically investigate the importance of each module in EmotionNAS, including the NAS-based approach, two-stream framework, and ISM fusion strategy. Finally, we visualize the confusion matrices and search results.

4.1. Comparison with Existing Works

To verify the effectiveness of our method, we treat current advanced approaches as our baselines. These methods include existing manually-designed and NAS-based models with consistent data division. We report baseline results according to their original papers. Experimental results in Table 1 show that EmotionNAS outperforms most existing methods and achieves a performance improvement of 2.8% over manually-designed models [9]. Compared with these approaches, our NAS-based framework can optimize structures and find more effective models than manually-designed approaches, automatically.

Meanwhile, we observe that EmotionNAS outperforms UniformNAS [32], a NAS-based approach, by 12.2% on UA and 1.2% on WA. But limited by handcrafted features, this baseline cannot leverage the useful information from deep features. Our framework, by contrast, not only enables efficient network design but also effectively fuses handcrafted and deep features, bringing better performance on SER.

4.2. Advantage of NAS

In this section, we compare the performance of several NAS-based and manually-designed models. Experimental results are shown in Table 2~3. For the spectrogram branch, NAS-based methods can achieve better performance with fewer parameters (see Table 2). For the wav2vec branch, our method also outperforms RNN and LSTM in emotion recognition (see Table 3). These results reveal the advantage of NAS in model design.

4.3. Necessity of Two Branches

In this section, we further reveal the necessity of two branches. As shown in Table 4, spectrogram and wav2vec achieve 57.3% and 66.2% on UA, 63.2% and 70.3% on WA, respectively. After multi-branch fusion, we further improve UA to 69.1% and

Table 5: *Importance of ISM.*

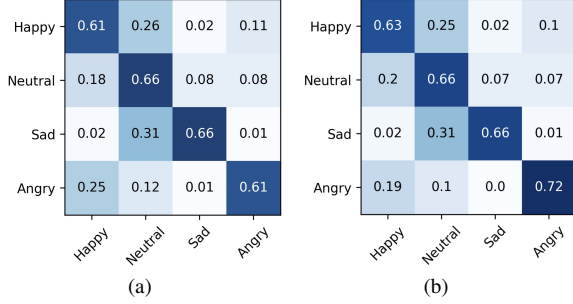| Method	UA(%)	WA(%)
Dominant branch (wav2vec)	66.2	70.3
Concat	67.1	68.5
Sum	63.8	66.0
Max	65.0	67.5
Min	68.1	65.2
EmotionNAS (Ours)	**69.1**	**72.1**

Figure 3: *Visualization of confusion matrices. (a) Results of* **Min***. (b) Results of EmotionNAS.*

WA to 72.1%. These results confirm that handcrafted and deep features contain complementary information in emotion recognition. Through the fusion process, we can achieve better classification performance on SER.

4.4. Effectiveness of ISM

To verify the effectiveness of our method, we compare ISM with other fusion strategies. Followed with the symbols in Section 2.3, we denote the outputs of two branches as X_1 and X_2. Here, X_1 is the dominant feature and H_1 is the linear transformation of X_1. **Concat** concatenates X_1 and X_2. **Sum**, **Max** and **Min** denote the corresponding operations on H_1 and X_2.

Experimental results are listed in Table 5. Among the four comparison strategies, only **Min** and **Concat** surpass the performance of the dominant branch. It indicates that most strategies cannot effectively fuse complementary information in different features. But through ISM, we can significantly improve the performance of UA and WA. These results demonstrate the effectiveness of our fusion strategy.

4.5. Visualization of Confusion Matrices

Figure 3 shows the confusion matrices for **Min** and EmotionNAS. We observe that EmotionNAS achieves better performance in most emotion categories, especially *happy* and *angry*. Our findings also highlight the limitations of traditional fusion strategies. Compared with **Min**, our fusion method ISM can effectively fuse multiple features and achieve better performance.

4.6. Visualization of Search Results

In this section, we further visualize the search results for each branch. Figure 4(a)∼4(b) show the search results for the spectrogram branch. There are two types of cells: the normal cell and the reduction cell. For cell k, the initial nodes consist of the outputs in previous cells, i.e., c_{k-2} and c_{k-1}. Figure 4(c) shows the search result for the wav2vec branch. For timestep t, the initial nodes consist of the input vector x_t and two hidden
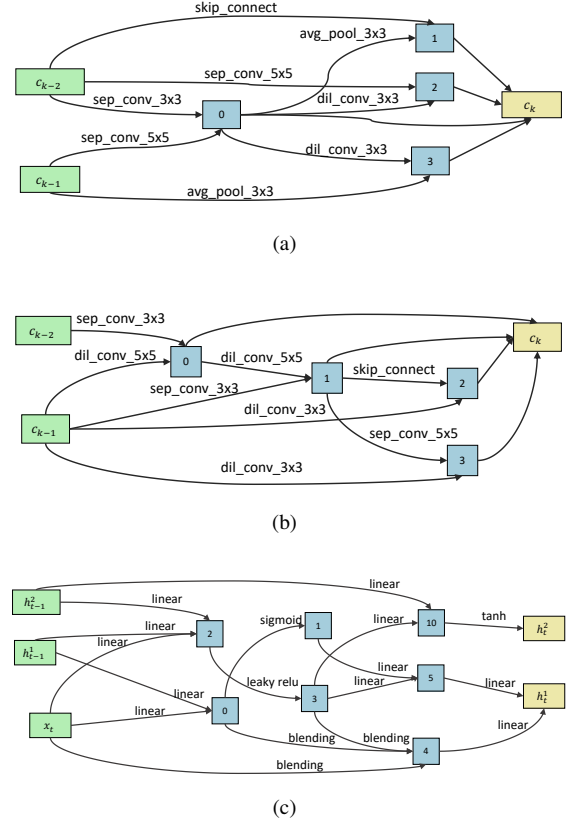

Figure 4: *Visualization of search results: (a) and (b) are search results of the normal cell and the reduction cell for the spectrogram branch; (c) is the search result for the wav2vec branch.*

states h_{t-1}^1 and h_{t-1}^2. We aim to generate new hidden states for the next timestep, i.e., h_t^1 and h_t^2.

5. Conclusions

In this paper, we propose a novel framework for SER, called "EmotionNAS". It takes spectrogram and wav2vec as the inputs, followed by NAS to optimize the network structure automatically. We further design ISM to integrate complementary information in different branches. Experimental results demonstrate that our method outperforms existing manually-designed and NAS-based methods. Meanwhile, we also systemically prove the importance of each component in EmotionNAS, including the NAS-based method, the two-stream framework and the ISM fusion strategy.

6. Acknowledgments

7. References

[1] M. Neumann and N. T. Vu, "Improving speech emotion recognition with unsupervised representation learning on unlabeled speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2019, pp. 7390–7394.

[2] H. Jian, L. Ya, T. Jianhua, L. Zheng, N. Mingyue, and Y. Jiangyan, "Speech emotion recognition using semi-supervised learning with ladder networks," in *Proceedings of the First Asian Conference on Affective Computing and Intelligent Interaction*, 2018, pp. 1–5.

[3] Z. Lian, Y. Li, J. Tao, and J. Huang, "Speech emotion recognition via contrastive loss under siamese networks," *CoRR*, vol. abs/1910.11174, 2019.

[4] M. M. H. E. Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit.*, vol. 44, no. 3, pp. 572–587, 2011.

[5] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech emotion recognition using deep learning techniques: A review," *IEEE Access*, vol. 7, pp. 117 327–117 345, 2019.

[6] X. Ma, Z. Wu, J. Jia, M. Xu, H. Meng, and L. Cai, "Emotion recognition from variable-length speech segments using deep learning on spectrograms," in *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association*, 2018, pp. 3683–3687.

[7] W. Han, H. Ruan, X. Chen, Z. Wang, H. Li, and B. W. Schuller, "Towards temporal modelling of categorical speech emotion recognition," in *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association*, 2018, pp. 932–936.

[8] X. Wu, S. Liu, Y. Cao, X. Li, J. Yu, D. Dai, X. Ma, S. Hu, Z. Wu, X. Liu *et al.*, "Speech emotion recognition using capsule networks," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6695–6699.

[9] Z. Zhao, Q. Li, Z. Zhang, N. Cummins, H. Wang, J. Tao, and B. W. Schuller, "Combining a parallel 2d cnn with a self-attention dilated residual network for ctc-based discrete speech emotion recognition," *Neural Networks*, vol. 141, pp. 52–60, 2021.

[10] Y. Xu, L. Xie, X. Zhang, X. Chen, G.-J. Qi, Q. Tian, and H. Xiong, "Pc-darts: Partial channel connections for memory-efficient architecture search," in *International Conference on Learning Representations (ICLR)*, 2019, pp. 1–13.

[11] H. Pham, M. Guan, B. Zoph, Q. Le, and J. Dean, "Efficient neural architecture search via parameters sharing," in *International conference on machine learning*. PMLR, 2018, pp. 4095–4104.

[12] T. Elsken, J. H. Metzen, and F. Hutter, "Neural architecture search: A survey," *The Journal of Machine Learning Research*, vol. 20, no. 1, pp. 1997–2017, 2019.

[13] X. Chen, L. Xie, J. Wu, and Q. Tian, "Progressive differentiable architecture search: Bridging the depth gap between search and evaluation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1294–1303.

[14] B. Zoph and Q. Le, "Neural architecture search with reinforcement learning," in *International Conference on Learning Representations (ICLR)*, 2017, pp. 1–16.

[15] E. Real, S. Moore, A. Selle, S. Saxena, Y. L. Suematsu, J. Tan, Q. V. Le, and A. Kurakin, "Large-scale evolution of image classifiers," in *International Conference on Machine Learning*. PMLR, 2017, pp. 2902–2911.

[16] H. Liu, K. Simonyan, and Y. Yang, "Darts: Differentiable architecture search," in *International Conference on Learning Representations (ICLR)*, 2019, pp. 1–13.

[17] M. Li, Q. Jiang, H. Lin, and H. An, "Gdarts: A gpu-based runtime system for dataflow task programming on dependency applications," in *IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking*, 2019, pp. 547–552.

[18] X. Chen, L. Xie, J. Wu, and Q. Tian, "Progressive differentiable architecture search: Bridging the depth gap between search and evaluation," in *2019 IEEE/CVF International Conference on Computer Vision, ICCV*, 2019, pp. 1294–1303.

[19] C. Ying, A. Klein, E. Christiansen, E. Real, K. Murphy, and F. Hutter, "Nas-bench-101: Towards reproducible neural architecture search," in *International Conference on Machine Learning*. PMLR, 2019, pp. 7105–7114.

[20] N. Klyuchnikov, I. Trofimov, E. Artemova, M. Salnikov, M. Fedorov, and E. Burnaev, "Nas-bench-nlp: Neural architecture search benchmark for natural language processing," *CoRR*, vol. abs/2006.07116, 2020.

[21] C. Vinola and K. Vimaladevi, "A survey on human emotion recognition approaches, databases and applications," *ELCVIA: electronic letters on computer vision and image analysis*, pp. 24–44, 2015.

[22] S. Khorram, Z. Aldeneh, D. Dimitriadis, M. G. McInnis, and E. M. Provost, "Capturing long-term temporal dependencies with convolutional networks for continuous emotion recognition," in *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association*, 2017, pp. 1253–1257.

[23] P. Li, Y. Song, I. McLoughlin, W. Guo, and L. Dai, "An attention pooling based representation learning method for speech emotion recognition," in *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association*, 2018, pp. 3087–3091.

[24] Y. Song, J. Liu, L. Wang, R. Yu, and J. Dang, "Multi-stage graph representation learning for dialogue-level speech emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6432–6436.

[25] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," in *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association*, 2019, pp. 3465–3469.

[26] S. Siriwardhana, T. Kaluarachchi, M. Billinghurst, and S. Nanayakkara, "Multimodal emotion recognition with transformer-based self supervised feature fusion," *IEEE Access*, vol. 8, pp. 176 274–176 285, 2020.

[27] Z. Lian, B. Liu, and J. Tao, "Ctnet: Conversational transformer network for emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 985–1000, 2021.

[28] 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, pp. 335–359, 2008.

[29] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8. Citeseer, 2015, pp. 18–25.

[30] M. Marcus, B. Santorini, and M. A. Marcinkiewicz, "Building a large annotated corpus of english: The penn treebank," *Computational Linguistics*, vol. 19, no. 2, pp. 313–330, 1993.

[31] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[32] X. Wu, S. Hu, Z. Wu, X. Liu, and H. Meng, "Neural architecture search for speech emotion recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6902–6906.