Interspeech 2024
1-5 September 2024, Kos, Greece

E-ODN: An Emotion Open Deep Network for Generalised and Adaptive Speech Emotion Recognition

Liuxian Ma[1,2,3,#], *Lin Shen*[1,2,#], *Ruobing Li*[1,2], *Haojie Zhang*[1,2], *Kun Qian*[1,2,*], *Bin Hu*[1,2,*], *Björn W. Schuller*[4,5], *and Yoshiharu Yamamoto*[6]

[1]Key Laboratory of Brain Health Intelligent Evaluation and Intervention (Beijing Institute of Technology), Ministry of Education, Beijing, 100081, China
[2]School of Medical Technology, Beijing Institute of Technology, Beijing, 100081, China
[3]School of Information and Electronics, Beijing Institute of Technology, Beijing, 100081, China
[4]GLAM – the Group on Language, Audio, & Music, Imperial College London, UK
[5]CHI – Chair of Health Informatics, MRI, Technical University of Munich, Germany
[6]Educational Physiology Laboratory, Graduate School of Education, The University of Tokyo, Japan
{qian, bh}@bit.edu.cn

Abstract

Recognising the widest range of emotions possible is a major challenge in the task of Speech Emotion Recognition (SER), especially for complex and mixed emotions. However, due to the limited number of emotional types and uneven distribution of data within existing datasets, current SER models are typically trained and used in a narrow range of emotional types. In this paper, we propose the Emotion Open Deep Network (E-ODN) model to address this issue. Besides, we introduce a novel Open-Set Recognition method that maps sample emotional features into a three-dimensional emotional space. The method can infer unknown emotions and initialise new type weights, enabling the model to dynamically learn and infer emerging emotional types. The empirical results show that our recognition model outperforms the state-of-the-art (SOTA) models in dealing with multi-type unbalanced data, and it can also perform finer-grained emotion recognition.

Index Terms: Speech Emotion Recognition, Open-set Recognition, Dynamic Learning, Three-dimensional Emotional Space

1. Introduction

Speech Emotion Recognition (SER) is a crucial component of natural human-computer interaction [1]. Speech emotion recognition is defined as extracting the emotional state of a speaker through his or her speech [2, 3]. A good quality emotional database is essential for SER [4, 5, 6]. However, due to the high costs of collecting emotional speech databases and annotation challenges, most databases primarily contain neutral emotions while containing a small portion of happiness and sadness, and even less coverage of emotions such as social emotions, causing a severe imbalance in the datasets [7, 8]. Thus, it is necessary to consider such imbalance in a dataset when attempting to extract emotional representations.

A popular approach to address data imbalance is by using data augmentation techniques [9, 10]. Using generative models such as Generative Adversarial Networks (GANs) to synthesise samples for augmenting training data is a workable approach. For instance, researchers in [11], [12], and [13] have enhanced SER performance on imbalanced datasets based on GANs. However, the authors of [14] have argued that when generating the synthetic feature vectors to augment SER systems, the lack of sufficiently large labelled datasets can lead to convergence issues in vanilla GANs. The authors of [15] utilise diffusion models for data augmentation in SER, achieving higher-quality generated samples compared to GANs.

Mixup [16] is also a common method that blends the features and labels of two different samples to generate new samples with new labels. The authors of [14] have employed "mixup" to train GANs for synthesising emotional feature generation as well as learning compressed emotional representations. Additionally, Meng *et al.* [17] have also implemented the mixup strategy to enhance speech recognition performance on small datasets. However, in speech, mixup has a significant drawback. Kang *et al.* [18] have argued that creating samples via mixup results in generated speech samples where two speakers engage in conversation, rather than speech samples similar to both speakers.

Consequently, although current methods can address imbalances in the current dataset and improve model performance, they cannot accurately classify emotions that are not covered (or poorly covered) by the corpus, nor can it learn new categories on the fly during the classification process. To address this issue, this paper proposes an open-set recognition model called E-ODN, which is based on emotional space mapping. This model not only recognises known data but also identifies unknown data and infers the underlying new emotions. In addition, we propose an emotional space mapping method that maps sample acoustic features into the PAD three-dimensional emotional model described below to obtain their emotion coordinate "feature vectors", enabling label inference for unknown categories. Furthermore, based on the distances between emotional vectors in the PAD space, we introduce a method to initialise new weights, allowing the model to learn newly added emotional categories more quickly and effectively preventing overfitting.

In summary the contributions of this paper are:

- We introduce an open set recognition strategy to address data imbalance in SER. Our strategy enables the model to identify and reason about unknown emotions, resulting in superior performance compared to state-of-the-art (SOTA) methods.
- We propose a model that can transform discrete emotional labels into continuous three-dimensional emotional vectors.

This work was partially supported by the National Key R&D Program of China (Nos. 2023YFC2506804 and 2022YFC3500503), the National Natural Science Foundation of China (Nos. 62272044, and 62227807), the Ministry of Science and Technology of the People's Republic of China with the STI2030-Major Projects (Nos. 2021ZD0201900, and 2021ZD0200601), and the Teli Young Fellow Program from the Beijing Institute of Technology, China. (*Liuxian Ma and Lin Shen contributed equally to this work. Corresponding authors: Kun Qian and Bin Hu.*)

10.21437/Interspeech.2024-685

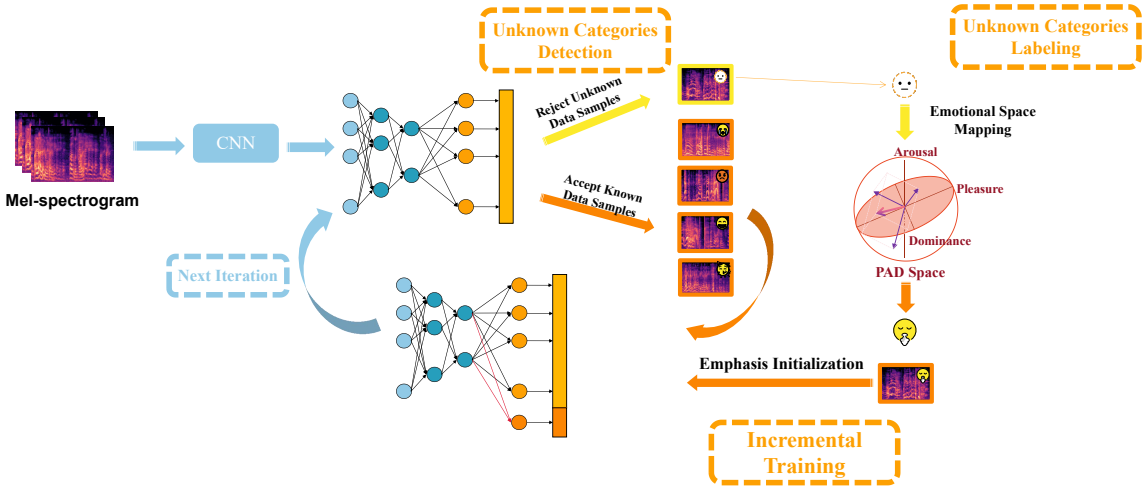

Figure 1: *The workflow of Emotion Open Deep Network. The entire workflow is primarily divided into three parts: Unknown Category Detection, Unknown Category Labelling, and Incremental Training.*

This model is endowed with the capability of fine-grained prediction for various emotions.

- We suggest that our emotional space mapping method is able to replace the manual annotation step required in open-set models.

2. Related Works

2.1. Emotional Space Representation

Mehrabian *et al.* [19, 20] find in their research that emotional experience can be evaluated based on three independent dimensions: Pleasure (P), Arousal (A), and Dominance (D). Based upon this, they proposed the PAD model of emotion. The PAD emotional space is a kind of emotional quantification method that can be used to describe the emotional state in the three-dimensional space allowing to obtain the emotional similarity and difference between each emotional state by spatial distance. Any coordinate position $\boldsymbol{e}(P, A, D)$ in the space is mapped to the corresponding emotional state E.

2.2. Open-Set Recognition Methods

The fundamental baseline model [21], trained with the softmax cross-entropy loss, has demonstrated reasonable effectiveness in both closed-set classification and unknown class detection. To enhance the detection mechanism for unknowns, OpenMax [22] introduces probabilistic modifications to the softmax activation, drawing upon extreme value theory [23]. DOC [24], alternatively, replaces the softmax cross-entropy with one-vs-rest logistic regression, finding its utility in rejecting invalid topics in natural language processing. RPL [25] proposes maximising inter-class separation in a reciprocal manner, followed by a variant that employes synthetically generated adversarial unknown classes. CPN [26] learns embedding metrics by modelling each known class as a group of multiple prototypes. PROSER [27] leverages latent mixup samples [28] as a synthetic unknown class, positioning their representations proximally to known class representations. Another approach employs multiple one-vs-rest networks [29] to address over-confidence and poor generalisation, relying on a collective decision score for effective open-set recognition.

Recently, it was demonstrated that the basic SCE [30] baseline could surpass other baselines if trained with robust data augmentation and SOTA optimisation techniques. However, [31] shows that a prior emphasis on well-separated discriminative embeddings remains crucial for effective open-set recognition.

3. Methods

3.1. Unknown Category Detection

Figure 1 displays the overall architecture of our proposed model. For the model, the primary task is to accept known samples while rejecting unknown ones, ensuring minimal risk in open spaces. Yu *et al.* [32] proposed a multi-class triplet thresholding method to detect unknown categories. Its core lies in calculating the triplet threshold for each category, namely the acceptance threshold η, rejection threshold μ, and distance rejection threshold δ. The triplet threshold for category i is calculated as:

$$\eta_i = \frac{1}{X_i} \sum_{j=1}^{X_i} F_{i,j} \tag{1}$$

$$\mu_i = \epsilon * \eta_i \tag{2}$$

$$\delta_i = \rho * \frac{1}{X_i} \sum_{j=1}^{X_i} (F_{i,j} - S_{i,j}), \tag{3}$$

where $F_{i,j}$ and $S_{i,j}$ represent the highest and second highest confidence values, j is the correctly classified sample of category i, and ϵ and ρ are the adjustable factors. The symbol X_i denotes the total number of correctly classified samples within the set $\mathcal{X}_i$, which corresponds to category i.

A data sample is classified as belonging into category l when the index of its highest confidence value p_l matches l and the value exceeds η_l. In contrast, a sample is considered unknown if all the confidence values are below μ. For values falling between η and μ, δ is utilised to assist in detecting unknown categories within the challenging samples. If the distance metric indicated by δ is sufficiently large, we accept the data sample as belonging to category label l.

3.2. Unknown Category Labelling

In the traditional ODN model, manual labelling is required for unknown categories. However, due to the finiteness and interconnectedness of emotions, we can infer the specific type of unknown emotion based on its relationship with known emotions.

Table 1: *Partial Standard Emotional Vector in the PAD Three-Dimensional Emotional Space.*

Emotion	Standard Emotional Vector
Happy	(2.77, 1.21, 1.42)
Optimistic	(2.48, 1.05, 1.75)
Relaxed	(2.19, -0.66, 1.05)
Surprise	(1.72, 1.71, 0.22)
Mild	(1.57, -0.79, 0.38)
Boring	(-0.53, -1.25, -0.84)
Sad	(-0.89, 0.17, -0.70)
Fearful	(-0.93, 1.30, -0.64)
Anxiety	(-0.95, 0.32, -0.63)
Disgusted	(-1.80, 0.40, 0.67)
Anger	(-1.98, 1.10, 0.60)
Animosity	(-2.08, 1.00, 1.12)

Thus, in our method, we adopt the emotional space mapping method to automatically infer emotions.

The emotional space mapping method involves mapping the unknown emotions samples onto a three-dimensional PAD emotional space, resulting in a feature vector f_i. Subsequently, the distance between f_i and the existing emotional standard vectors is calculated to assess their similarity.

Assuming there are n known classes, the computation method for the feature vector $\boldsymbol{f}_i$ of X_i is:

$$\boldsymbol{f_i} = \sum_{j=1}^{X_i} p_{ij} * \boldsymbol{F}_j \quad j = 1, 2, \ldots, n, \tag{4}$$

where p_{ij} represents the confidence value associated with the known j^{th} category and $\boldsymbol{F}_j$ denotes the standard emotional vector for the known j^{th} category. According to the research of Mehrabian *et al.* [19], the standard emotion vectors in PAD space are shown in Table 1. The distances d_j between f_i and each of the standard emotional vectors is measured as:

$$d_j = |\boldsymbol{f}_i - \boldsymbol{F}_j| \quad j = 1, 2, \ldots, m \quad (m > n), \tag{5}$$

where m is the number of standard emotional vectors and $|*|$ represents the modulo operation on $*$.

The index of the closest standard vector is chosen as the estimated label for the sample. We do not need to deliberately remove the known emotional standard vectors because the triplet thresholding method we employ for unknown type detection ensures that the emotional feature vector f_i for the unknown class X_i maintains a significant distance from the known emotional vectors in the PAD space. This ensures that the estimated label for X_i is unlikely to match any of the known emotion categories. Furthermore, in the next section, when updating the model weights, we utilise the distances between f_i and the known types F_j.

3.3. Incremental Training

After inferring the emotional labels of unknown samples, the new categories need to be incorporated into training to update the network. However, due to the scarcity of unknown data, training from scratch can lead to issues such as overfitting, resulting in the model unstable. Therefore, we use the distances between the new samples and known emotions in the PAD emotional space, as well as the existing model weights, to initialise the weights for the new categories.

$$w_{n+1} = \frac{1}{n} \sum_{j=1}^{n} d'_j * w_j, \tag{6}$$

where w_j represents the weight column of the j^{th} category, and d'_j denotes the value of d_j after undergoing mean normalisation.

The detected unknown data is then utilised to fine-tune the model. We adjust the learning rates differently for known and unknown classes, allowing the new weights to learn at a faster rate. Following this phase of learning, our model is equipped with the ability to recognise unknown emotions.

4. Experiments and Results

In this section, we select datasets and baseline models to empirically evaluate the effectiveness of three modules within our proposed model: unknown category recognition, positional category inference, and open emotion recognition.

4.1. Datasets

We use the USC IEMOCAP [33] database for the experiments. The dataset comprises 12 hours of English audiovisual data organised into five session. Each utterance in the database is annotated by at least three evaluators with categorical emotion labels from: happy, sad, neutral, angry, surprised, exited, frustrated, disgusted, fearful, or other. For emotions that are more pronounced (such as anger and excitement), there are over 1 000 samples for each emotion. However, for emotions that are more difficult to express (such as fear and surprise), the number of samples is below 200. The IEMOCAP dataset effectively illustrates the characteristic of uneven data distribution commonly found in speech emotion datasets. This uneven distribution precisely represents the challenge our model aims to address.

In the experiment, we select five emotions (happy, sad, frustrated, angry, excited) as known categories for training data and three emotions (surprised, disgusted, fearful) as the unknown category data. For other baseline models without open recognition capabilities, we conduct unified training and testing using eight emotional datasets.

We train and evaluate our model by using a leave-one-session-out cross-validation approach [34]. In this strategy, four sessions are designated for training, with the remaining session serving as the test set. This process was repeated five times, ensuring that all sessions are used for both training and testing. At each training epoch, we assess the model's performance on the test set, and the final reported results reflect the average scores across these five iterations.

4.2. Implementation Details

We extract the Mel spectrogram of the audio samples as the input for the model and implement the Convolutional Neural Network (CNN) using ResNet-50 [35] with pre-trained initial weights.

For model training, we adopt a similar approach to the ODN [32] model. In the first stage, we provide known data samples to train a recognized classifier. After the training is completed, the system use the activation values from the trained model to detect unknown categories. In the second stage, we incrementally add unknown samples for further training.

During the testing phase, apart from evaluating the classification results for known and unknown categories, we also compare the model's regression results by calculating the difference between the inferred and actual values of the emotional space coordinates for each sample.

4.3. Baselines and Evaluation Protocols

For the unknown type detection stage, we choose OpenMax [22] and Open Long-Tailed Recognition (OLTR) [36] models as baselines to compare their abilities in distinguishing unknown types.

Table 2: *Unknown Detection Results of E-ODN Measured by F1-Score(in [%]).*

Methods	F-score
OSDN	59.3
OLTR	70.5
E-ODN	**71.1**

Table 3: *CCC Scores (in [%]) during Unknown Category Labelling.*

Methods	$CCC_{avg} \uparrow$	$CCC_A \uparrow$	$CCC_D \uparrow$	$CCC_P \uparrow$
preCPC	47.5	**61.6**	30.2	48.6
E-ODN	**48.1**	55.5	**35.8**	**50.8**

In the stage of labelling unknown types, which essentially involves regressing the emotional P, A, D values, we select the pre Contrastive Predictive Coding (preCPC) [37] model as the baseline and evaluate it using the Concordance Correlation Coefficient (CCC). For the overall speech emotion recognition model, we compare and assess it against several SOTA models, including GLAM [38], DRN-MHSA [39], and GAN-SER [13], using weighted accuracy (WA), unweighted accuracy (UA), and Micro-F1 as metrics. The baseline model employs the same emotional categories and training/testing split as our model. However, unlike our model, other models do not require a two-stage training process. Instead, they directly perform a classification task.

4.4. Results and Discussions

4.4.1. Unknown Classes Detection

To evaluate the model's effectiveness in detecting unknown classes, we assigned all known class labels as 0 and unknown class labels as 1, and then calculated the model's classification F-score. As shown in Table 2, the E-ODN model performs better in handling unknown classes compared to the baselines. Specifically, it improves by 11.8 % compared to the OSDN model based on OpenMax, indicating that the multi-class triplet thresholding method is more suitable for open-set detection of emotions than OpenMax.

4.4.2. Unknowns Labelling

When inferring unknown emotion labels, we project the samples into the PAD emotional space using existing features, effectively obtaining three dimensional values that represent the emotional aspects of the sample. We calculate the CCC value for our model via leveraging the original dimensional labels from the dataset. According to Table 3, our model demonstrates comparable regression accuracy to specialised regression models. This finding not only verifies the high degree of correctness in our label predictions but also indicates that our model can adapt to different fine-grained emotion recognition tasks. It can operate with discrete labels as classification outcomes or continuous values as recognition outputs, providing flexibility and adaptability to various scenarios.

It is worth noting that our model's performance in estimating the arousal dimension is not as good as some existing models. We believe this is due to the relatively similar arousal values in our training data, which results in insufficient generalisation capability for newly introduced emotions.

Table 4: *Performance Comparison between E-ODN and its Counterparts in Terms of WA (%), UA (%), and Micro-F1 (%).*

Methods	WA	UA	Micro-F1
GLAM	56.3	55.2	55.7
DRN-MHSA	55.1	53.5	54.3
GAN-SER	61.1	60.2	60.7
E-ODN	**61.2**	**63.1**	**62.2**

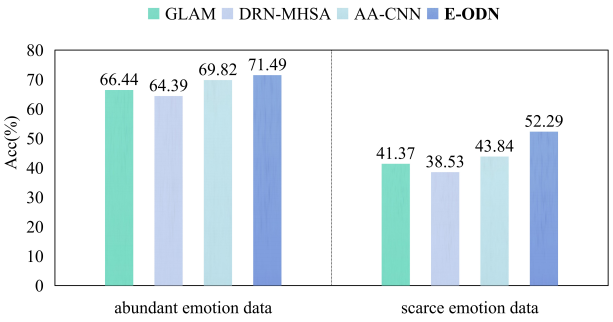

Figure 2: *The accuracy of data classification based on different data volume types.*

4.4.3. SER with Unbalanced Data

The overall results of the SER model are presented in Table 4. Our model exhibits superior performance compared to the baseline in handling unbalanced datasets, particularly in terms of the under-represented class (UA) metric. Our model achieve a score of 63.1 %, which is an improvement of 2.9 % compared to the previous best model.

To further assess the model's ability to discriminate between emotional categories represented by both abundant and scarce data samples, we categorise the existing data types into two groups: abundant and scarce, based on their respective sample sizes. Within each group, we calculate the average recognition accuracy for each emotional category. Finally, we compare these averages visually, as shown in the Figure 2.

For emotions with abundant data, the scores of our model and SOTA models are generally high. Besides, the differences among them are relatively small. However, in the scenario of emotions with scarce data, our model performs considerably higher accuracy compared to the SOTA models. This suggests that our model exhibits stability in handling imbalanced data scenarios as well as suffering no considerable drops in accuracy due to the lack of individual data points. The consistent performance across different emotional categories, regardless of data availability, further underscores the robustness and reliability of our model.

5. Conclusions

In this work, we proposed E-ODN, an open-set SER model, to address the issue of imbalanced data distributions in SER datasets. This model can determine unknown emotion classes and infer these new emotions. Furthermore, our model realised incorporating incremental learning as well as reducing the model's dependency on new samples. The results of experiments based on the IEMOCAP dataset demonstrated that our model exhibits enhanced robustness compared to existing models when dealing with imbalanced data distributions. In future works, we aim to apply our SER model to the task of emotional speech generation, aiming to produce speech data with richer emotional expressions.

6. References

[1] A. Triantafyllopoulos, B. W. Schuller, G. İymen, M. Sezgin, X. He, Z. Yang, P. Tzirakis, S. Liu, S. Mertes, E. André *et al.*, "An overview of affective speech synthesis and conversion in the deep learning era," *Proceedings of the IEEE*, pp. 1–24, 2023.

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

[3] J. Tavernor, M. Perez, and E. M. Provost, "Episodic memory for domain-adaptable, robust speech emotion recognition," in *Proc. INTERSPEECH*, 2023, pp. 1–5.

[4] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech Communication*, vol. 48, no. 9, pp. 1162–1181, 2006.

[5] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach, "Emotional speech: Towards a new generation of databases," *Speech Communication*, vol. 40, no. 1-2, pp. 33–60, 2003.

[6] Y. Lee and T. Kim, "Robust and fine-grained prosody control of end-to-end speech synthesis," in *Proc. ICASSP*, 2018, pp. 5911–5915.

[7] Z. Tu, B. Liu, W. Zhao, R. Yan, and Y. Zou, "A feature fusion model with data augmentation for speech emotion recognition," *Applied Sciences*, pp. 1–18, 2023.

[8] S. Wang, J. Guðnason, and D. Borth, "Learning emotional representations from imbalanced speech data for speech emotion recognition and emotional text-to-speech," *arXiv preprint arXiv:2306.05709*, pp. 1–5, 2023.

[9] D. A. Van Dyk and X.-L. Meng, "The art of data augmentation," *Journal of Computational and Graphical Statistics*, vol. 10, no. 1, pp. 1–50, 2001.

[10] T. Koike, K. Qian, B. W. Schuller, and Y. Yamamoto, "Transferring cross-corpus knowledge: An investigation on data augmentation for heart sound classification," in *Proc. EMBC*, 2021, pp. 1976–1979.

[11] L. Yi and M.-W. Mak, "Adversarial data augmentation network for speech emotion recognition," in *Proc. APSIPA ASC*. IEEE, 2019, pp. 529–534.

[12] A. Chatziagapi, G. Paraskevopoulos, D. Sgouropoulos, G. Pantazopoulos, M. Nikandrou, T. Giannakopoulos, A. Katsamanis, A. Potamianos, and S. Narayanan, "Data augmentation using gans for speech emotion recognition." in *Proc. INTERSPEECH*, 2019, pp. 171–175.

[13] S. Wang, H. Hemati, J. Guðnason, and D. Borth, "Generative data augmentation guided by triplet loss for speech emotion recognition," *arXiv preprint arXiv:2208.04994*, pp. 1–5, 2022.

[14] S. Latif, M. Asim, R. Rana, S. Khalifa, R. Jurdak, and B. W. Schuller, "Augmenting generative adversarial networks for speech emotion recognition," *arXiv preprint arXiv:2005.08447*, pp. 1–5, 2020.

[15] M. I. Malik, S. Latif, R. Jurdak, and B. W. Schuller, "A Preliminary Study on Augmenting Speech Emotion Recognition using a Diffusion Model," in *Proc. INTERSPEECH*, 2023, pp. 646–650.

[16] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, pp. 1–13, 2017.

[17] L. Meng, J. Xu, X. Tan, J. Wang, T. Qin, and B. Xu, "Mixspeech: Data augmentation for low-resource automatic speech recognition," in *Proc.ICASSP*, 2021, pp. 7008–7012.

[18] W. H. Kang, J. Alam, and A. Fathan, "L-mix: A latent-level instance mixup regularization for robust self-supervised speaker representation learning," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1263–1272, 2022.

[19] A. Mehrabian and J. A. Russell, "A measure of arousal seeking tendency," *Environment and Behavior*, vol. 5, no. 3, p. 315, 1973.

[20] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *Journal of Research in Personality*, vol. 11, no. 3, pp. 273–294, 1977.

[21] L. Neal, M. Olson, X. Fern, W.-K. Wong, and F. Li, "Open set learning with counterfactual images," in *Proc. ECCV*, 2018, pp. 613–628.

[22] A. Bendale and T. E. Boult, "Towards open set deep networks," in *Proc. CVPR*, 2016, pp. 1563–1572.

[23] W. J. Scheirer, A. Rocha, R. J. Micheals, and T. E. Boult, "Meta-recognition: The theory and practice of recognition score analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1689–1695, 2011.

[24] L. Shu, H. Xu, and B. Liu, "Doc: Deep open classification of text documents," in *Proc. EMNLP*, 2017, pp. 2911–2916.

[25] G. Chen, L. Qiao, Y. Shi, P. Peng, J. Li, T. Huang, S. Pu, and Y. Tian, "Learning open set network with discriminative reciprocal points," in *Proc. ECCV*. Springer, 2020, pp. 507–522.

[26] H.-M. Yang, X.-Y. Zhang, F. Yin, Q. Yang, and C.-L. Liu, "Convolutional prototype network for open set recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 5, pp. 2358–2370, 2020.

[27] D.-W. Zhou, H.-J. Ye, and D.-C. Zhan, "Learning placeholders for open-set recognition," in *Proc. CVPR*, 2021, pp. 4401–4410.

[28] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *Proc. ICLR*, 2018, pp. 1–13.

[29] J. Jang and C. O. Kim, "Collective decision of one-vs-rest networks for open-set recognition," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–8, 2022.

[30] S. Vaze, K. Han, A. Vedaldi, and A. Zisserman, "Open-set recognition: A good closed-set classifier is all you need," in *Proc. ICLR*, 2021, pp. 1–27.

[31] T. Kasarla, G. Burghouts, M. van Spengler, E. van der Pol, R. Cucchiara, and P. Mettes, "Maximum class separation as inductive bias in one matrix," *Advances in Neural Information Processing Systems*, vol. 35, pp. 19 553–19 566, 2022.

[32] Y. Shu, Y. Shi, Y. Wang, Y. Zou, Q. Yuan, and Y. Tian, "Odn: Opening the deep network for open-set action recognition," in *Proc. ICME*, 2018, pp. 1–6.

[33] 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.

[34] C. Ding, J. Li, D. Zong, B. Li, T.-H. Zhang, and Q. Zhou, "Stable Speech Emotion Recognition with Head-k-Pooling Loss," in *Proc. INTERSPEECH*, 2023, pp. 661–665.

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.

[36] Z. Liu, Z. Miao, X. Zhan, J. Wang, B. Gong, and S. X. Yu, "Large-scale long-tailed recognition in an open world," in *Proc. CVPR*, 2019, pp. 2537–2546.

[37] M. Li, B. Yang, J. Levy, A. Stolcke, V. Rozgic, S. Matsoukas, C. Papayiannis, D. Bone, and C. Wang, "Contrastive unsupervised learning for speech emotion recognition," in *Proc. ICASSP*, 2021, pp. 6329–6333.

[38] W. Zhu and X. Li, "Speech emotion recognition with global-aware fusion on multi-scale feature representation," in *Proc. ICASSP*, 2022, pp. 6437–6441.

[39] R. Li, Z. Wu, J. Jia, S. Zhao, and H. Meng, "Dilated residual network with multi-head self-attention for speech emotion recognition," in *Proc. ICASSP*, 2019, pp. 6675–6679.