*INTERSPEECH 2021*
30 August – 3 September, 2021, Brno, Czechia

# Metric Learning Based Feature Representation with Gated Fusion Model for Speech Emotion Recognition

*Yuan Gao*[1,†], *JiaXing Liu*[1,†], *Longbiao Wang*[1,2,*], *Jianwu Dang*[1,2,3]

[1]Tianjin Key Laboratory of Cognitive Computing and Application,
College of Intelligence and Computing, Tianjin University, Tianjin, China
[2]TianjinUniversity-HuiyanTechnology Joint AI Lab, TianjinUniversity, Tianjin, China
[3]Japan Advanced Institute of Science and Technology, Ishikawa, Japan
{yuan_gao, jiaxingliu, longbiao_wang}@tju.edu.cn, jdang@jaist.ac.jp

## Abstract

Due to the lack of sufficient speech emotional data, the recognition performance of existing speech emotion recognition (SER) approaches is relatively low and requires further improvement to meet the needs of real-life applications. For the problem of data scarcity, an increasingly popular solution is to transfer emotional information through pre-training models and extract additional features. However, the feature representation needs further compression because the training object of unsupervised learning is to reconstruct input, making the latent representation contain non-affective information. In this paper, we introduce deep metric learning to constrain the feature distribution of the pre-training model. Specifically, we propose a triplet loss to modify the representation extraction model as a pseudo-siamese network and achieve more efficient knowledge transfer for emotion recognition. Furthermore, we propose a gated fusion method to learn the connection of features extracted from the pre-training model and supervised feature extraction model. We conduct experiments on the common benchmarking dataset IEMOCAP to verify the performance of the proposed model. The experimental results demonstrate the advantages of our model, outperforming the unsupervised transfer learning system by 3.7% and 3.88% in weighted accuracy and unweighted accuracy, respectively.

**Index Terms**: speech emotion recognition, human-computer interaction, metric learning, gated fusion model

## 1. Introduction

With the rapid development of artificial intelligence, communication between humans and computers has become pervasive in our daily lives. Speech emotion recognition aims to identify human emotions from speech automatically, a robust SER system can benefit many applications such as call center service and psychological disease diagnosis. In the past two decades, previous works presented that deep learning models such as the combination of a convolutional neural network (CNN) and long short-term memory (LSTM) can extract emotional features without expert knowledge and significantly improve the performance of the SER system [1, 2, 3]. To avoid overfitting, the supervised learning approaches need a large quantity of labelled data, which is difficult to collect in this field [4]. Available emotional speech datasets are relatively small compared to other speech processing tasks such as automatic speech recognition.

Despite the difficulty in collecting sufficient speech data with high-quality emotional labels in a natural environment, there are still potential solutions to deal with data scarcity. Some researchers conduct experiments on several existing speech datasets which have similar emotion annotations (i.e., cross-corpus speech emotion recognition) [5, 6, 7]. Nevertheless, various factors in different datasets such as speaker and recording environment remain challenges for building a robust cross-corpus SER system. Other researchers pre-train the model and transfer emotion information from these datasets to assist with the emotion recognition task [8]. As autoencoders [9, 10] are emphasized for feature modelling, this is the most common model for knowledge transfer. However, the objective of this structure is to retain as much information as possible from the input, so the bottleneck feature also contains various non-affective information [11]. To deal with the spontaneity of unsupervised learning, we incorporate deep metric learning (DML) as a constraint for the pre-training model.

DML can learn an embedding space for input pairs or triplets to generate discriminative features with a relatively smaller intra-class distance and larger inter-class distance. This training strategy has attracted increasing attention and is widely used in various tasks such as few-shot learning [12] and person re-identification [13, 14]. Inspired by the success in visual analysis, researchers have introduced metric learning to improve the performance of the SER system. Lian et al. [15] proposed two kinds of contrastive loss for SER. The authors conduct experiments using different feature selection methods to evaluate the model performance. However, siamese networks focus only on inputs with the same emotion category. Aiming to obtain the largest separation between positive and negative pairs, Huang et al. [16] utilize hard-positive mining techniques to evaluate the performance of triplet loss. In [17], Dai et al. employ softmax cross-entropy loss and center loss together to cooperately learn discriminative features from variable length spectrograms. To improve the robustness of cross-corpus SER, Zhang et al. [18] proposed a f-Similarity Preservation Loss function and conducted experiments using soft labels.

Previous approaches mainly incorporate metric learning to improve the feature discriminability of supervised learning approaches. In this study, to address the data scarcity in SER, we pre-train an unsupervised convolutional autoencoder to transfer prior knowledge. A triplet loss is designed to add an additional constraint between the pre-training model and the deep CNN architecture. The proposed loss function enables the autoencoder to share the same feature distribution with CNN model. Another difficulty for emotion recognition is how to preserve the complementary information of features learned from different models. Common feature fusion approaches such as concatenating the feature vector [19] or feeding the features into dif-

† equal contribution

* corresponding author

http://dx.doi.org/10.21437/Interspeech.2021-1133

ferent classifiers and assembling the decision [20] may ignore the connections between these feature representations. To address this problem, we proposed a gated fusion model to learn the contribution of features extracted from the deep CNN model and pre-training model. We also incorporate attention mechanism [21] to model the relative dependencies in different parts of the input sequence.

The rest of this paper is organized as follows. We describe the network details in Section 2. The emotional datasets used in this work are presented in Section 3. In Section 4, experiment results are provided to evaluate the effectiveness of our model. This paper is concluded in Section 5.

## 2. Proposed pseudo-siamese network

Since the commonly used deep learning classifiers only focus on finding a decision boundary to separate different types of emotions, we use DML to improve the feature discriminability for the SER task. Moreover, the feature representations learned by the pre-trained autoencoder and CNN are fused through our proposed gated fusion model.

### 2.1. Metric learning for knowledge transfer

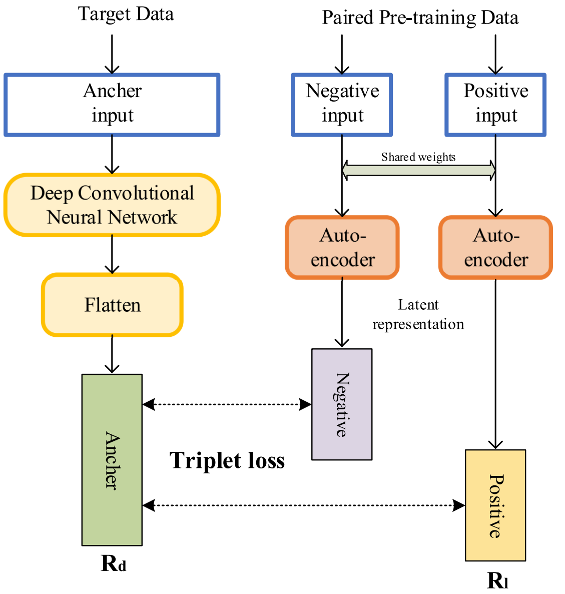


Figure 1: *Network structure of the proposed pseudo-siamese feature extractor*

As shown in Figure 1, the proposed feature extraction model contains two main components: a pre-trained autoencoder for knowledge transfer; and a deep CNN for feature extraction. Previous studies have shown that pre-trained autoencoders can efficiently extract additional features from a large quantity of speech data P(X), which can be used to assist the emotion recognition of target data. The objective function of the traditional autoencoder can be defined as:

$$L_{AE} = \text{argmin}\,||P(X) - P'(X)||^2 \tag{1}$$

where P(X) is the data used for pre-training. Although the unsupervised pre-training model can transfer the emotion information from P(X), due to the lack of supervised information filtering, the bottleneck feature still contains noise information for SER. In this paper, we introduce metric learning to learn a distance metric for triplet inputs to measure their similarities and encourage the pre-training model to extract more discriminative features for emotion recognition. We use a triplet loss to combine the deep CNN model and the pre-trained autoencoder as a pseudo-siamese network. Triplet loss is proposed in FaceNet [22] and calculated on the triplet input data ($X$, $X_p$, $X_n$), where $X$ and $X_p$ share the same emotion labels and X and $X_n$ have different emotion labels. The input $X$ is usually defined as the anchor of input triplets. As shown in Figure 2, this training strategy learns an embedding space and uses margin $M$ to make the feature distributions learned from the same category closer to each other than those learned from different categories. To achieve knowledge transfer and extract additional features, the proposed pseudo-siamese network is trained in the following two phases: (1) the pre-training phase and (2) the feature extraction phase.

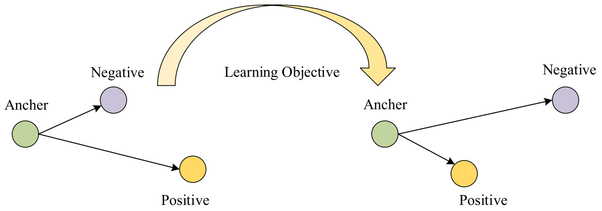


Figure 2: *The learning objective of triplet loss*

In the pre-training phase, we use the convolutional autoencoder to transfer prior knowledge by reconstructing P(X). The input of the CNN is target data X. We define X as the anchor input of the siamese network. To generate the triplet input, we divide the P(X) into positive pairs and negative pairs, respectively. As the emotion annotations of X and P(X) are inconsistent, in this specific task, the triplet input is grouped according to their valence and arousal information. Our proposed triplet loss function is defined as follows:

$$pos\_dist = ||f(x_i^a) - f(x_i^p)||_2 \tag{2}$$

$$neg\_dist = ||f(x_i^a) - f(x_i^n)||_2 \tag{3}$$

$$L_{tri} = max(post\_dist - neg\_dist + M, 0) \tag{4}$$

where $pos\_dist$ and $neg\_dist$ are the Euclidean distances between features learned from positive pairs and negative pairs. $L_{tri}$ represents the proposed triplet loss. Combining these two models as pseudo-siamese networks, the pre-training model is constrained by both $L_{AE}$ and the triplet loss and therefore shares the same feature distribution with the supervised CNN model. The objective function of the feature extraction model can be defined as:

$$L = \alpha L_{cro} + \beta L_{tri} + \gamma L_{AE} \tag{5}$$

where $L_{cro}$ is the cross-entropy loss function of the deep CNN model. $\alpha$, $\beta$, and $\gamma$ are trade-off parameters to control the weight of each loss term. In the pre-training phase, $\alpha$ and $\beta$ are set as 0.3, and $\gamma$ is set as 0.4 to achieve better knowledge transfer. To avoid the influence of P(X) on the CNN model, we load only the weights and bias of the autoencoder and fine-tune the framework using target data.

In the feature extraction phase, target data are directly fed into the CNN and one branch of the autoencoder to extract the bottleneck features. During feature extraction, $\alpha$ is set as 0.6, while $\beta$ and $\gamma$ are set as 0.2. Through our proposed triplet loss,

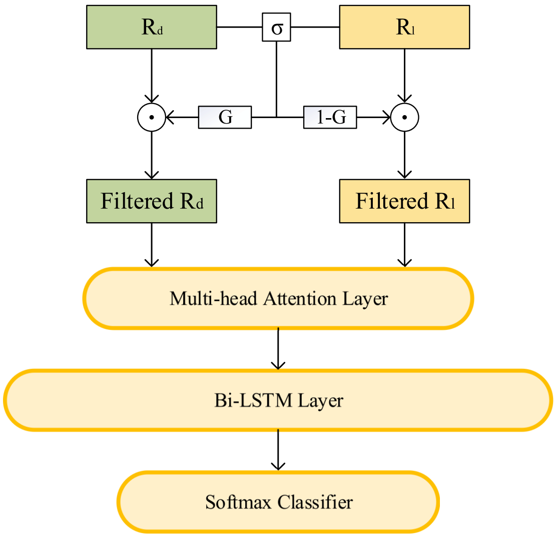


Figure 3: *Network structure of the gated fusion model*

the latent representation $R_l$ can not only transfer prior knowledge from P(X) but also share the benefit of supervised learning. Moreover, the deep representation $R_d$ is extracted by the deep CNN architecture, for which the parameters are the same as those of Satt et al. [23].

### 2.2. Gated fusion model

For the aforementioned two kinds of feature representations $R_d$ and $R_l$, we designed a gated fusion model to obtain more effective feature representations for each input utterance. As shown in Figure 3, our proposed fusion model can be described as:

$$G = \sigma(W_g[R_d, R_l] + b_g) \tag{6}$$

$$F_d = f(G \odot R_d) \tag{7}$$

$$F_l = f((1 - G) \odot R_l) \tag{8}$$

where $W_g$ and $b_g$ are trainable parameters and bias. $\sigma$ is the sigmoid activation function to learn the contributions of two kinds of input features. G denotes the gate vector, for which the value ranges from 0 to 1. In Equations (7) to (8), $f$ is the activation function and $\odot$ represents an elementary product. Our gate mechanism controls their contributions by multiplying the corresponding input features and produces the filtered representations $F_d$ and $F_l$. Furthermore, an attention mechanism is introduced to capture the salient emotional parts and reduce the information loss during the feature fusion stage. We fed the filtered $R_d$ and $R_l$ to an attention layer to obtain the input of the emotion classifier. Through this gated feature fusion model, we can reduce the irrelevant information for emotion recognition and learn the connection between $R_d$ and $R_l$. Given the obtained representations of each utterance, we used Bi-LSTM followed with a softmax layer to obtain the emotion output.

## 3. Experimental setup

### 3.1. Corpora description

We used the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [24] to evaluate our model. This dataset contains 12 hours of audio, video, facial motion information, and textual transcriptions. Both scripted and improvised audio data are used in this work. Following the common practice [25, 26, 27], we choose four emotions for our experiments (namely, happy, sad, angry, and neutral).

Table 1: *Emotion mapping for triplet input*

| **Valence** | | | | |
|---|---|---|---|---|
| Corpus | IEMOCAP | EmoV-DB | SAVEE | EMODB |
| Positive | Happy, Neutral | Amused, Neutral | Surprise, Neutral, Joy | Happy, Neutral |
| Negative | Angry, Sad | Disgust, Angry | Disgust, Angry Sad, Fear | Disgust, Angry, Sad, Fear |
| **Arousal** | | | | |
| Corpus | IEMOCAP | EmoV-DB | SAVEE | EMODB |
| High | Happy, Angry | Amused, Angry | Surprise, Angry, Joy, Fear | Happy, Angry, Fear |
| Low | Neutral, Sad | Disgust, Neutral | Disgust, Neutral, Sad, Fear | Disgust, Neutral, Sad, Fear |

To achieve knowledge transfer, we choose three public emotional datasets to pre-train the autoencoder: Surrey audio-visual expressed emotion (SAVEE) [28], Berlin Emotional Speech Database (EMO-DB) [29], and Emotional Voices Database (EmoV-DB) [30]. The SAVEE dataset contains audio-visual recordings of four male subjects. This dataset includes 480 native English utterances: 60 for each of six basic emotions [31] and 120 utterances for neutral. The EMO-DB dataset was performed by ten professional actors in a recording environment. The actors were asked to express each sentence in seven emotional states. This dataset comprises 535 utterances. EmoV-DB contains 7,000 samples of four male speakers. Four classes of emotions are used in this paper, namely amused, disgust, angry and natural. We choose these three emotion corpora because they all contain sufficient emotion information and are available to the community.

### 3.2. Experimental settings

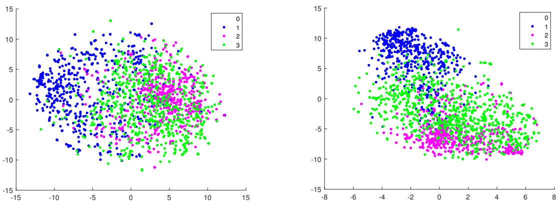
To achieve knowledge transfer from different kinds of corpora, we divide the pre-training data into positive and negative pairs as the input of the siamese network. Due to the aforementioned inconsistent emotional annotations of the different corpora, we use the valence and arousal information as a unified standard. As shown in Table 1, we reference Schuller et al. [4] to map categorical emotions to binary valence and arousal. Utterances in P(X) that share the same valence and arousal with anchor data are grouped as positive pairs. Similarly, those which have different valence or arousal are grouped as negative pairs.

We use categorical labels for the emotion recognition task. The input signals of both X and P(X) are sampled at 16 kHz. Given an utterance, we split it into 265-ms segments. The input spectrogram is calculated for each segment, with a frame size of 25 ms. Therefore, the time × frequency of the input spectrogram is 32 × 129, and the batch size is set as 128. We use adadelta as the optimizer. In this study, we randomly choose 80% of the data in IEMOCAP for training and 20% for testing.

## 4. Results and analysis

### 4.1. Visualization analysis

To evaluate the impact of metric learning, we introduce t-distributed stochastic neighbor embedding (t-SNE) [32] to visualize the feature distribution of the pre-training model.



(a) autoencoder (b) proposed model

Figure 4: *The t-SNE visualization for the pre-training model.*

As shown in Figure 4, the blue (angry) points perform well in both representations. The traditional autoencoder exhibits poor performance for the separation of green (sad) and magenta (happy) points. In the feature distribution of the proposed model, this phenomenon is mitigated. Moreover, the magenta points and blue points are closer to each other. We use white points to obfuscate the feature distribution of neutral, as neutral is not well aggregated in both plots. The same situation is also reported in [8].

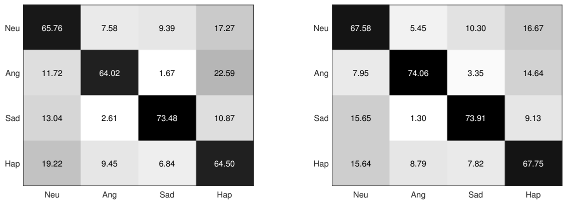
### 4.2. Classification performance of the proposed method

We choose weighted accuracy (WA) and unweighted accuracy (UA) as evaluation criteria in this work. The classification results of five comparative experiments are presented in Table 2 to validate the effectiveness of the proposed model. In the CNN-BLSTM baseline system, only target data X is used. The pre-training (PT) model contains a pre-trained convolutional autoencoder to extract additional representations from P(X). The features extracted from the CNN and autoencoder are concatenated as the input of the BLSTM classifier. In the foundation of the PT model, PT_G further introduces the proposed gated model for feature fusion. The pseudo-siamese network (PSN) utilizes the triplet loss to combine the deep CNN model and pre-trained autoencoder. PSN_G is the proposed pseudo-siamese network with gated fusion model for feature compression.

Table 2: *Classification results of comparative experiments*

| model | WA(%) | UA(%) |
|---|---|---|
| CNN-BLSTM | 63.12 | 63.47 |
| PT | 65.38 | 65.95 |
| PT_G | 66.64 | 66.94 |
| PSN | 68.89 | 69.27 |
| PSN_G | 70.34 | 70.82 |

The experimental results for the baseline system and PT model have shown that a pre-trained autoencoder can transfer emotion information and improve the performance of an emotion recognition system, which supports the conclusion of previous studies [8, 33]. To verify the impact of metric learning, we compare the performances of PSN and PT. As shown in Table 2, the PSN model outperforms the PT model, with absolute increases of 3.51% and 3.32% in WA and UA, respectively. Compared to the PT_G model, our proposed PSN_G also achieves an increase of more than 3% in both WA and UA. This result indicates that the pre-training model can share the benefits of supervised learning through the proposed triplet loss. Furthermore, both PT_G and PSN_G achieve better performance than the corresponding model using only concatenation for feature compression, which reveals the effectiveness of our proposed gated fusion model. To gain more insights into the experimental results, we also analyse the confusion matrices of PT_G and PSN_G in Figure 5.



|  | Neu | Ang | Sad | Hap |
|---|---|---|---|---|
| Neu | 65.76 | 7.58 | 9.39 | 17.27 |
| Ang | 11.72 | 64.02 | 1.67 | 22.59 |
| Sad | 13.04 | 2.61 | 73.48 | 10.87 |
| Hap | 19.22 | 9.45 | 6.84 | 64.50 |

(a) PT_G

|  | Neu | Ang | Sad | Hap |
|---|---|---|---|---|
| Neu | 67.58 | 5.45 | 10.30 | 16.67 |
| Ang | 7.95 | 74.06 | 3.35 | 14.64 |
| Sad | 15.65 | 1.30 | 73.91 | 9.13 |
| Hap | 15.64 | 8.79 | 7.82 | 67.75 |

(b) PSN_G

Figure 5: *The confusion matrices.*

In the confusion matrix of PT_G, 22.59% of angry utterances are misclassified as happy. Both happy and angry have relatively higher arousal. Hence, the classifier needs to distinguish these two emotions through their valence. Our proposed approach decreased this error recognition rate by 7.95% and significantly improved the recognition accuracy for angry by 10.04%. Compared with that in the PT_G model, neutral and happy also achieve steady increases of more than 1.82%. The improvement of our proposed model can be attributed to the proposed triplet loss, which compensates for the weaknesses of unsupervised transfer learning.

## 5. Conclusions and future work

In this paper, we incorporated deep metric learning into the feature extraction model as a pseudo-siamese network to improve the performance of the SER system. Experiments on IEMOCAP demonstrated the effectiveness of the proposed method. Compared with the traditional pre-training model, our proposed model achieved an average performance of 70.34% and 70.82% for WA and UA, respectively, with absolute improvements of 3.7% and 3.88% over previous works. Furthermore, we present the visualization analysis for high-level representation obtained from both the proposed and common pre-training models. The feature distribution has better convergence with the help of triplet loss. In the future, we will focus on the cross-corpus SER to further evaluate the generalization ability of metric learning.

## 6. Acknowledgements

This work was supported by the National Key R&D Program of China under Grant 2018YFB1305200, the National Natural Science Foundation of China under Grant 61771333 and the Tianjin Municipal Science and Technology Project under Grant 18ZXZNGX00330.

# 7. References

[1] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2227–2231.

[2] L. Guo, L. Wang, J. Dang, L. Zhang, and H. Guan, "A feature fusion method based on extreme learning machine for speech emotion recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 2666–2670.

[3] P. Tzirakis, J. Zhang, and B. W. Schuller, "End-to-end speech emotion recognition using deep neural networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5089–5093.

[4] B. Schuller, B. Vlasenko, F. Eyben, M. Wöllmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll, "Cross-corpus acoustic emotion recognition: Variances and strategies," *IEEE Transactions on Affective Computing*, vol. 1, no. 2, pp. 119–131, 2010.

[5] B. Zhang, E. M. Provost, and G. Essl, "Cross-corpus acoustic emotion recognition with multi-task learning: Seeking common ground while preserving differences," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 85–99, 2017.

[6] H. Kaya and A. A. Karpov, "Efficient and effective strategies for cross-corpus acoustic emotion recognition," *Neurocomputing*, vol. 275, pp. 1028–1034, 2018.

[7] J. Gideon, M. McInnis, and E. M. Provost, "Improving cross-corpus speech emotion recognition with adversarial discriminative domain generalization (addog)," *IEEE Transactions on Affective Computing*, 2019.

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

[9] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.

[10] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 1096–1103.

[11] J. Deng, Z. Zhang, F. Eyben, and B. Schuller, "Autoencoder-based unsupervised domain adaptation for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 21, no. 9, pp. 1068–1072, 2014.

[12] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1199–1208.

[13] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Deep metric learning for person re-identification," in *2014 22nd International Conference on Pattern Recognition*. IEEE, 2014, pp. 34–39.

[14] M. Hirzer, P. M. Roth, and H. Bischof, "Person re-identification by efficient impostor-based metric learning," in *2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance*. IEEE, 2012, pp. 203–208.

[15] Z. Lian, Y. Li, J. Tao, and J. Huang, "Speech emotion recognition via contrastive loss under siamese networks," in *Proceedings of the Joint Workshop of the 4th Workshop on Affective Social Multimedia Computing and First Multi-Modal Affective Computing of Large-Scale Multimedia Data*, 2018, pp. 21–26.

[16] J. Huang, Y. Li, J. Tao, Z. Lian *et al.*, "Speech emotion recognition from variable-length inputs with triplet loss function." in *Interspeech*, 2018, pp. 3673–3677.

[17] D. Dai, Z. Wu, R. Li, X. Wu, J. Jia, and H. Meng, "Learning discriminative features from spectrograms using center loss for speech emotion recognition," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7405–7409.

[18] B. Zhang, Y. Kong, G. Essl, and E. M. Provost, "f-similarity preservation loss for soft labels: A demonstration on cross-corpus speech emotion recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 5725–5732.

[19] S.-L. Yeh, Y.-S. Lin, and C.-C. Lee, "An interaction-aware attention network for speech emotion recognition in spoken dialogs," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6685–6689.

[20] E. M. Albornoz and D. H. Milone, "Emotion recognition in never-seen languages using a novel ensemble method with emotion profiles," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 43–53, 2015.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[22] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[23] A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms." in *Interspeech*, 2017, pp. 1089–1093.

[24] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, p. 335, 2008.

[25] P. Yenigalla, A. Kumar, S. Tripathi, C. Singh, S. Kar, and J. Vepa, "Speech emotion recognition using spectrogram & phoneme embedding." in *Interspeech*, 2018, pp. 3688–3692.

[26] F. Bao, M. Neumann, and N. T. Vu, "Cyclegan-based emotion style transfer as data augmentation for speech emotion recognition." in *Interspeech*, 2019, pp. 2828–2832.

[27] A. Ando, R. Masumura, H. Kamiyama, S. Kobashikawa, and Y. Aono, "Speech emotion recognition based on multi-label emotion existence model." in *Interspeech*, 2019, pp. 2818–2822.

[28] P. Jackson and S. Haq, "Surrey audio-visual expressed emotion (savee) database," *University of Surrey: Guildford, UK*, 2014.

[29] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[30] A. Adigwe, N. Tits, K. E. Haddad, S. Ostadabbas, and T. Dutoit, "The emotional voices database: Towards controlling the emotion dimension in voice generation systems," *arXiv preprint arXiv:1806.09514*, 2018.

[31] P. Ekman, W. V. Friesen, M. O'sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W. A. LeCompte, T. Pitcairn, P. E. Ricci-Bitti *et al.*, "Universals and cultural differences in the judgments of facial expressions of emotion." *Journal of personality and social psychology*, vol. 53, no. 4, p. 712, 1987.

[32] L. Van Der Maaten, "Learning a parametric embedding by preserving local structure," in *Artificial Intelligence and Statistics*. PMLR, 2009, pp. 384–391.

[33] S. Ghosh, E. Laksana, L.-P. Morency, and S. Scherer, "Representation learning for speech emotion recognition." in *Interspeech*, 2016, pp. 3603–3607.