Unsupervised Cross-Corpus Speech Emotion Recognition Using a Multi-Source Cycle-GAN

Bo-Hao Su, *Student Member, IEEE,* and Chi-Chun Lee, *Senior Member, IEEE*

Abstract—Speech emotion recognition (SER) plays a crucial role in understanding user feelings when developing artificial intelligence services. However, the data mismatch and label distortion between the training (source) set and the testing (target) set significantly degrade the performances when developing the SER systems. Additionally, most emotion-related speech datasets are highly contextualized and limited in size. The manual annotation cost is often too high leading to an active investigation of unsupervised cross-corpus SER techniques. In this paper, we propose a framework in unsupervised cross-corpus emotion recognition using multi-source corpus in a data augmentation manner. We introduced Corpus-Aware Emotional CycleGAN (CAEmoCyGAN) including a corpus-aware attention mechanism to aggregate each source datasets to generate the synthetic target sample. We choose the widely used speech emotion corpora the IEMOCAP and the VAM as sources and the MSP-Podcast as the target. By generating synthetic target-aware samples to augment source datasets and by directly training on this augmented dataset, our proposed multi-source target-aware augmentation method outperforms other baseline models in activation and valence classification.

Index Terms—Speech emotion recognition, data augmentation, cross corpus, unsupervised learning, multi-sources attention

1 Introduction

As advanced deep learning algorithms and hardware capacity have developed, artificial intelligence (AI) services and products have proliferated in recent years. Speech emotion recognition (SER) techniques have become more prevalent and continuously been applied on the real-world applications, e.g., satisfaction measurement of customer [1], health care service [2], call centers [3], intelligent vehicle assistance [4], [5], [6] and human-machine interactions [7], [8], [9]. These diverse applications are ubiquitous in our daily life. Hence, developing robust algorithms plays a critical role to ensure a wide adoption of SER. Most speech emotion datasets are limited in scales and highly contextualized which is detrimental when training a model to be robust for cross-corpus scenarios. Applications suffer from adapting multiple unique training corpora to provide general emotion recognition; that is, severe domain mismatch across different datasets significantly hinders the model's generalization performance. To overcome the barrier caused by the discrepancy between source and target, instead of labeling an adequate amount of data from the target corpus, the most practical way is to transfer the knowledge from the existing labeled source corpus.

For many years, researchers have investigated algorithms including unsupervised domain adaptation (which learns to mitigate discrepancies by mapping source distribution to target distribution) and domain-invariant learning for SER (which learns a common representation subspace for both domains) to map a source corpus to an unlabeled target database. For domain adaptation, techniques include a conventional matrix factorization method [10] and an advanced method [11] where Albanie et al. introducing cross-modal distillation network that transfers from video to audio. In [12], Zong et al. imposed a regularization that alleviates domain mismatch and proposed domain adaptive least square regression (DaLSR), which uses both labeled source data and unlabeled target data as auxiliary data during training. Latif et al. [13] proposed deep belief networks (DBN) formed from a stack of Restricted Boltzmann Machines (RBMs) to train using a greedy layer-wise schema, which demonstrates that the model is able to learn discriminative long-range features. For domain-invariant learning, Zhang utilized different normalization schemes to understand which schema would be better for handling undesirable domain-specific factors [14], and Shuller in [15] broadly explored the effect of different normalizations among six standard SER databases and discussed the variances and the strategies in this field. Huang explored a common feature space and domain-specific feature space using PCANet [16], as well as several transfer learning methods [17]. Most literature addresses the unsupervised cross corpus emotion recognition in these two major directions, and the current state-of-the-art method is based on domain adversarial learning model [18], [19].

In addition to the feature domain mismatch, emotion labels can also become distorted across different datasets. Chao et al. [20] pointed out that the distortion in emotion semantics should be considered even in similar collection settings, and Wei [21] stated that the subdomains (e.g., categories and classes) should be jointly considered while training a transfer learning model. This type of research indicates the matching among feature representations does not intrinsically equal to aligning emotion labels between corpora. For SER corpora, in particular, the size is limited and somewhat restricted by the scenarios, which implies the variability coverage may be insufficient. Under such circumstances, even when matching the feature representation of source and target domains, which usually assumes

B.-H. Su and C.-C. Lee are with the Department of Electrical Engineering, National Tsing Hua University, Hsinchu, Taiwan 300044 e-mail: (borrissu@gapp.nthu.edu.tw, cclee@ee.nthu.edu.tw).

that the same category would then be aligned between two different domains, the model may not be robust and it also may misalign the emotion labels due to the peculiarity of the application setting. In fact, most of the prior works have focused on these two major issues when conducting unsupervised cross-corpus emotion recognition: 1) feature distribution mismatch, 2) emotion label distortion.

Aside from these two common issues, additionally, the known limited scale, and given that emotion manifestation is complex, it is unclear whether there is enough variability in the data collected to accomplish SER (even in within-database recognition), as evident in many recent augmentation works. Recent studies show that SER results would benefit from increasing the variability within the corpus using data augmentation methods [22], [23], [24], [25] that explicitly expand the diversity and increase the number of training samples. This approach has resulted in superior performance when compared to other state-of-the-art models for within-database emotion recognition. Thus, we argue that to properly handle unsupervised cross-corpus emotion recognition, one needs to simultaneously consider issues of domain/label mismatch and limited data variability. To this end, we have proposed a method in our previous work [26] that transfers emotional information from source to target by augmenting source corpus using three types of synthetic target-aware samples. This model addresses the former issues simultaneously and achieves state-of-the-art accuracy.

While a large body in the variants of works has been conducted in this field, most of the methods mentioned above focus on one-to-one mapping, which selects a specific source and target and transfers between these two corpora. Thus, choosing the **appropriate** pair of corpora that could better perform one-to-one mapping is significant, e.g., SER performance would be better when transferring the IEMOCAP to the MSP-IMPROV than when transferring the IEMOCAP to the CreativeIT due to the similarity in scenario setting. However, this constraint of choosing the **appropriate** corpora makes this type of research not scalable and inefficient, and it would further under-utilize the multiple existing and available labeled data sources. At the moment, there is no literature investigating how to handle this many-to-one unsupervised cross corpus emotion recognition situation.

In fact, we argue that many-to-one mapping is a necessary next step to advance unsupervised cross corpus emotion recognition. Imagine a scenario in reality, e.g., Source-A is similar to Target-C in content-wise, and Source-B is similar to Target-C in its environment. When the model directly trained with either Source-A or Source-B solely would not be the most suitable source model for the target by itself. However, if we can make use of Source-B, i.e., merging both samples with proper weights, the synthesized sample could better adapt to the target corpus. To this end, in this paper, we propose a corpus-aware emotional data augmentation method that handles the usage of multiple-source datasets. Specifically, we introduce a novel corpus-aware attention mechanism while training a cycleGAN-based model comprising a source-to-target generator, a target-to-source generator, two source discriminators, and a target discriminator.

The primary contribution of this work beyond our previous work [26] is that we integrate information from multiple sources to generate target-aware samples, implementing the concept of many-to-one mapping for cross-corpus emotion recognition. Specifically, we integrate the usefulness in the uniqueness of each source using our proposed corpus-aware attention mechanism and further conduct comprehensive analyses to understand our proposed framework and visualize the working mechanism behind our proposed corpus-aware attention mechanism. The attention mechanism refers to a target sample as a reference sample and fuses the samples from two different source datasets to form the most similar synthetic sample possible. After training the proposed architecture, we augment the source data by generating the synthetic target-aware samples. We can then train a recognizer on this augmented dataset, aggregation from different unique source datasets, to and directly infer on the target dataset.

We evaluate on the widely used speech emotion datasets, the IEMOCAP and the VAM, as our training sources and the MSP-Podcast as target dataset. The performance of our proposed method surpasses all the state-of-the-art augmentation strategies. We achieved a 61.64% unweighted average recall (UAR) for arousal and a 44.62% UAR for valence on MSP-Podcast in an unsupervised cross corpus SER setting, which outperforms the state-of-the-art model by 7.09% and 0.69%, respectively, in absolute points. The rest of the paper is organized as follows: section 2 lists and discusses related works. Our proposed architecture's components and objective functions are detailed in section 3. In section 4, we describe the experimental settings and the compared baseline models. Then, in section 5, we present the experimental results and analysis. The conclusion, section 6, summarizes the paper and the proposed method.

2 RELATED WORKS

Most SER models that are trained on one corpus usually fail to generalize to other datasets. The mismatch between speech emotion databases has become one of the biggest hurdles to generalize SER. Additionally, the complex and time-consuming nature of speech emotion data collection has limited the size of most datasets.

Researchers have developed sophisticated algorithms using various strategies to overcome these situations. Because of the aforementioned limitations, the recent SER research focuses on domain adaptation and domain-invariant learning algorithms that are essentially divided into three major directions: adversarial learning-based models, autoencoder-based models, and GAN-based models.

However, unlike the above methodologies, some studies tackle this problem using data augmentation methods, which increase the source domain's variability. In fact, this method has historically been successfully applied in automatic speech recognition (ASR) tasks [27], [28]. For its use in domain adaptation scenarios, Ko et al. [29] proposed simulating the reverberation speech data as a data augmentation way to solve the far-field problem in real-world environments. In [30], Hsu augmented the source data using a variational autoencoder, which learns a latent representation to transform the source to the target domain. To address the lack of public child speech data, which is also limited by

TABLE 1: Database Statistics

Corpus	Content	Data	Setting	# of utt	Aro. H/M/L(%)	Val. H/M/L(%)
IEMOCAP	English	Audio, Video, MOCAP	Scripted & Spontaneous	10039	16.85/72.07/11.08	19.39/48.5/32.11
VAM	German	Audio, Video, Faces	Spontaneous	947	19.53/68.22/12.25	1.06/69.9/29.04
MSP-Podcast(Test)	English	Audio	Spontaneous	9255	32.93/59.91/7.15	20.01/65.59/14.40

the age-dependent condition, Sheng [31] explored a GAN-based data augmentation method that increases the amount of data while simultaneously improving the recognition accuracy. The results demonstrate a large relative word error rate (WER) reduction of over 20%. Chen [32] proposed to increase data variation and diversity by combining the GAN-based model and multi-style training as a data augmentation method for ASR system, and the results show a 35% relative reduction in WER. Recently, a prevalent surge in the use of GAN-based generative architectures has led to promising performances on many modeling tasks. The cycle-consistent GAN [33] has demonstrated outstanding performance while conducting the image transformation between the source and target data.

In the following subsections, we elaborate on the four major methodologies, i.e., adversarial learning-based models, autoencoder-based models, GAN-based models, and data augmentation-based models.

2.1 Adversarial Learning-Based Models

Abdelwahab proposed DANN [19], which learns a representation between a source and a target using adversarial invariant learning and has demonstrated promising performance. Then, Gideon [18] integrated the concept of the WGAN [34] to build ADDoG, in which a critic module is imposed to force the representation to be similar to the two distributions. Xiao in [35] also modified the raw DANN model using a bottleneck fully connected layer to create the CGDANN, which includes a variational autoencoder-based feature extractor and considers the reconstruction loss, emotion loss, and gradient reverse on domain loss simultaneously.

2.2 Autoencoder-Based Models

Models based on autoencoders were widely used to handle the representation learning tasks. Deng [36] proposed a denoising autoencoder (DAE) that adapts to the target domain using a combination of adaptive DAE (A-DAE) and DAE to learn the general representation of both the source and target domain. Deng in [37] also proposed an universum autoencoder-based model to learn the unsupervised representation from labeled data and explored the prior knowledge from unlabeled data to improve the SER performance. Neumann [38] improved the cross-corpus SER accuracy by concatenating the features from RNN-based autoencoders and CNN-based models. Furthermore, Huang [16] applied principal components analysis (PCA) filters to extract the domain-invariant and domain-specific features that improved the performance of cross-corpus SER.

2.3 GAN-Based Models

Several papers have leveraged the advantages of the GAN-based model in the main architecture to adapt the representation from source to target domain. Hoffman et al. [39] proposed the cycle consistent adversarial domain adaptation (CyCADA), which considers the loss due to discrimination by the source or target while also taking the main task loss into account. Using a similar concept, the latest generative model, CycleEmotionGAN [40] was proposed to mitigate the mismatch between the source and target domain, and its structure ensures the emotion semantic consistency as well. Both architectures consider semantic consistency loss and aim to align the sentiment content between the source and target datasets.

2.4 Data Augmentation-Based Models

Leveraging the outstanding image transformation quality achieved by using cycleGAN, Bao [25] also demonstrated cycleGAN's superior performance for SER tasks by synthesizing fake samples for data augmentation. Chatziagapi et al. [22] proposed to adopt a balancing GAN (BAGAN) that generates synthetic spectrograms for the minority classes, then augments the source data to balance the distribution of each class. In our previous work, we proposed a conditional cycle emotion GAN (CCEmoGAN) [26] as a target-aware data augmenter for source datasets to address the cross-corpus SER modeling by generating fake target samples with an extra condition vector to control specific emotions.

Data augmentation, as an approach that improves SER performance, has gained more and more attention in recent years. Therefore, inspired by the augmentation methods, in this paper, we apply the generative model to synthesize target-aware samples; this approach strengthens the source domain model's generalization capacity when it is applied to source data. Refer to [18], the cross-corpus learning methods can generally be divided into four quadrants: generative models, discriminative models, domain generalization methods, and domain adaptation methods. To better focus on the generative component, we compare the proper baseline work to our model in our experiments, which are described in the later section.

3 RESEARCH METHODOLOGY

3.1 Speech Emotion Corpus

In this work, we considered the distinctiveness of the available datasets and decided to use the IEMOCAP and the VAM as our source datasets because they comprise different languages, scenarios, and collection settings. The MSP-Podcast was chosen as our target corpus because it contains real-world speech. In the following subsections, we elaborate on these three corpora in detail.

3.1.1 Source 1: USC's IEMOCAP

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset [41] is composed of audio and video emotion clips collected by researchers at the University of Southern California (USC). It contains a total of approximately 12

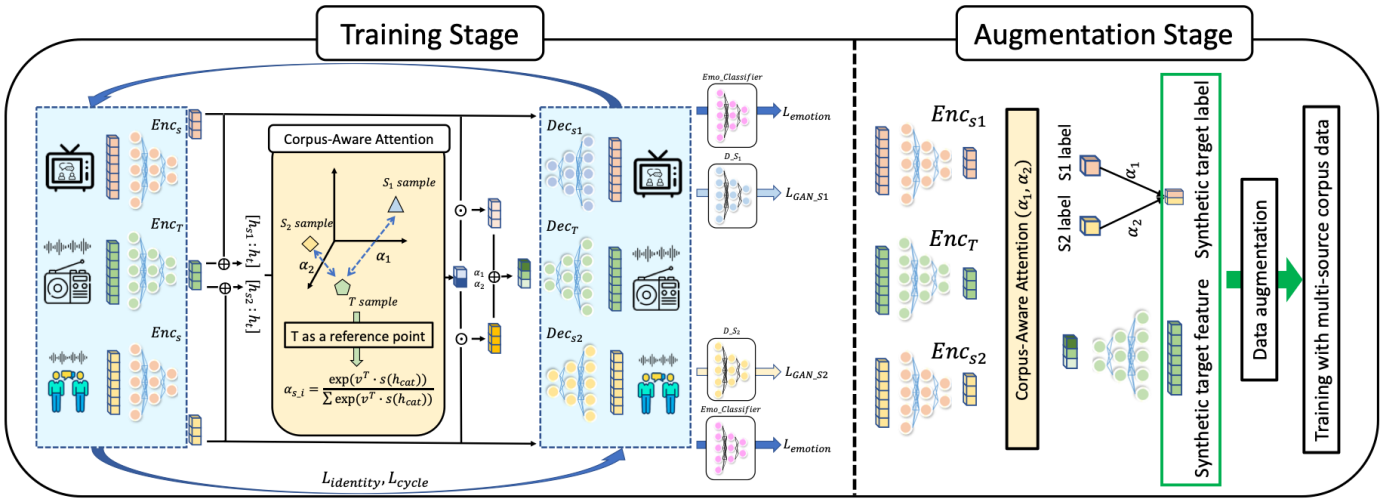


Fig. 1: Overview of the cross-corpus speech emotion recognition (SER) architecture using our proposed CAEmoCyGAN data augmentation. The training stage is used to train a generator for many-to-one mapping, and the augmentation stage augments the original source's data using target-aware samples to train a classifier for the target corpus.

hours of data that have been manually segmented into sentences. There are five dyadic sessions where each consists of two actors (one male and one female). Each session includes both spontaneous and scripted dialogue interactions. All utterances were annotated by at least three raters in terms of both dimensional attributes (arousal and valence) as well as categorical emotion labels. In this work, we used a total of 10039 sentences, and the arousal and valence label were divided into three classes using the boundary of [1, 2], (2, 4), [4, 5].

3.1.2 Source 2: Vera am Mittag (VAM)

The VAM corpus [42] is a spontaneous and emotionally-rich speech database collected from a German talk show. VAM collected audio data and facial expressions from a total of 47 speakers. Each broadcast consists of several multi-party dialogues (two to five people), and 70% of the speakers collected were 35 years old or younger at the time of collection. The corpus was extracted from a total of approximately 12 hours of data. The segmented sentences' annotations include the attributes of arousal, valence, and dominance. The annotation values range from -1 to 1. In this work, we used a total of 947 utterances annotated with both arousal and valence attributes. In addition, the utterances were divided into three classes according to the boundary [-1, -0.33], (-0.33, 0.33), [0.33, 1].

3.1.3 Target: MSP-Podcast

The MSP-Podcast [43] is a spontaneous speech emotion dataset collected from real-life podcasts. All the utterances are emotional or neutral speech from online podcasts and are segmented into durations of 2.75s to 11s. There are 33626 total utterances in the version used in this work and they are divided into three sets including a training set (19707 utterances), a validation set (4300 utterances), and a testing set (9255 utterances). Audio segments were annotated by Amazon Mechanical Turk (AMT) workers, who evaluated their perceptions of the utterances. At least five annotators rated the utterances using a scale ranging from 1 to 7, and the ratings concerned the arousal, valence, and dominance attributes. In this work, we used only 9255 utterances from the testing set as our target corpus. We adopted the boundary [1, 3), [3, 5), [5, 7] to split the emotion attributes into three classes.

3.2 Acoustic Features

We extracted 1582 dimensional functional features from each utterance using the openSMILE toolkit [44] with the Emobase config file. It is extracted by computing statistical functions on low-level descriptors (LLDs) such as pitch, energy, and Mel-Frequency Cepstrum Coefficients (MFCC). Moreover, there are two major reasons for using this feature set. The first one is the dimensionality, to stably converge the GAN training, neither too large nor few dimensions are favored in our architecture. Second, previous works have shown that the Emobase feature set can obtain competitive results when compared to eGeMAPs feature set [45]. Many recent works on SER [46], [47] also use this feature set as inputs. The min-max normalization schema is applied on a corpus-wise scale to fix the values ranging from -1 to 1, which improves the efficiency of the cycle-GAN training.

3.3 Corpus-Aware Emotional Cycle-GAN

The overview of our proposed method is presented in Fig. 1, where all the symbols and abbreviations are made consistent throughout the paper. The architecture is composed of a modified cycleGAN with a corpus-aware attention mechanism and the emotion consistency constraint. Conventionally, the cycleGAN is used for a one-to-one mapping between the source and target corpus.

However, instead of directly conducting one-to-one mapping, we imposed a corpus-aware attention mechanism on the middle stage of the cycle-GAN to integrate two latent representations from the IEMOCAP (S_1) and the VAM (S_2) using learnable attention weights (α_S). Using the corpus-aware attention mechanism, we achieved the **many-to-one** mapping between multiple sources (IEMOCAP-S_1, VAM-S_2) and the target (MSP_Podcast-T). In order to further preserve the emotion information, we additionally added an emotion consistency constraint to the reconstructed samples

from each source so they could be accurately classified. Finally, the well-trained CAEmoCyGAN was used to generate target-aware samples to augment the multi-source training corpus and re-train a classifier for the target corpus. In the following subsections, we detail each component and its corresponding loss function.

3.3.1 The CycleGAN

In this framework, we trained a bi-directional mapping function between the source and target corpus that contained two generators ($G_{S\to T}$, $G_{T\to S}$) and three corpus discriminators (D_{S_1}, D_{S_2}, D_T). Here, we use $G_{S\to T}$ as an encoder for the source (Enc_S) and a decoder for the target (Dec_T). The source encoder (Enc_S) maps samples from two different source domains to a common space, and the target decoder (Dec_T) maps the latent representation to the target domain. A common cycleGAN was applied as our base model here, and the standard GAN losses are defined as:

$$\mathcal{L}(G_{S\to T}, D_T) = \mathbb{E}_{T\sim P_{data}(T)}[\log D_T(T)] + \mathbb{E}_{S\sim P_{data}(S)}[\log(1 - D_T(Dec\_T(Enc\_S(S))))] \quad (1)$$

Unlike the normal cycleGAN, the GAN loss for the source was slightly modified in our framework. As for $G_{T\to S}$, we used two source decoders (Dec_S_i) where i corresponds to the source corpus index. The source GAN loss is defined as:

$$\mathcal{L}(G_{T\to S_i}, D_{S_i}) = \mathbb{E}_{S_i\sim P_{data}(S_i)}[\log D_{S_i}(S_i)] + \mathbb{E}_{T\sim P_{data}(T)}[\log(1 - D_{S_i}(Dec\_S_i(Enc\_T(T))))] \quad (2)$$

Therefore, the total GAN loss of our framework was:

$$\mathcal{L}_{GAN}(G_{T\to S_i}, G_{S\to T}, D_{S_i}, D_T) = \mathcal{L}(G_{T\to S_i}, D_{S_i}) + \mathcal{L}(G_{S\to T}, D_T) \quad (3)$$

To guarantee the stability of the training process and the reversibility of the generators, we also considered identity loss and cycle loss as well during training, and the two losses are defined as:

$$\mathcal{L}_{identity} = \mathbb{E}_{S_i\sim P_{S_i}}[||G_{T\to S_i}(S_i) - S_i||^2] + \mathbb{E}_{T\sim P_T}[||G_{S\to T}(T) - T||^2] \quad (4)$$

$$\mathcal{L}_{cycle} = \mathbb{E}_{S_i\sim P_{S_i}}[||G_{T\to S_i}(G_{S\to T}(S_i)) - S_i||^2] + \mathbb{E}_{T\sim P_T}[||G_{S\to T}(G_{T\to S_i}(T)) - T||^2] \quad (5)$$

3.3.2 The Corpus-Aware Attention Mechanism

The original cycleGAN architecture only conducted the one-to-one mapping between one source and one target dataset. When training a traditional cycleGAN, a pairing input like (S_1, T) or (S_2, T) is required. However, to achieve the many-to-one mapping in the cycleGAN, we introduced a corpus-aware attention mechanism in this work. While training our proposed network, the pairing inputs included two different source samples and one target sample (S_1, S_2, T); this combination forms a many-to-one mapping when input into the corpus-aware attention mechanism. According to our proposed method, the target samples were treated as reference points for both source samples, and attend on either source IEMOCAP or source VAM, whichever is more similar to the reference target point. First, we defined the h_{cat}, which is the concatenation of the hidden vector from generators' encoder, and the input to the corpus-aware attention mechanism is as follows:

$$h_{s_i} = Enc\_S(S_i) \quad (6)$$

$$h_{t_{ref}} = Enc\_T(T) \quad (7)$$

$$h_{cat} = (h_{s_1}||h_{t_{ref}})||(h_{s_2}||h_{t_{ref}})||...||(h_{s_i}||h_{t_{ref}}) \quad (8)$$

$$h_{cat} \in \mathbb{R}^{B\times C\times 2H} \quad (9)$$

where h_{s_i} is the hidden vector from Enc_S of $G_{S\to T}$, S_i is the sample from source corpus i, and $h_{t_{ref}}$ is the hidden vector from paired input of target sample and serves as a reference point in each batch from Enc_T of $G_{T\to S}$. B, C, and H represent the batch size, source corpus amount, and hidden dimension size of the encoder respectively.

The corpus-aware attention mechanism contains a dimension reduction layer and a trainable parameter, and is defined as:

$$s(h_{cat}) = tanh(Linear(h_{cat})), s(h_{cat}) \in \mathbb{R}^{B\times AH\times C} \quad (10)$$

$$\alpha_S = \frac{exp(v^T \cdot s(h_{cat}))}{\sum_{i=1}^{C} exp(v^T \cdot s(h_{cat}))} \quad (11)$$

$$\alpha_S \in \mathbb{R}^{B\times 1\times C} \quad (12)$$

where $s(\cdot)$ is a score function for the energy computation, h_{cat} is the input to the attention mechanism and represents the concatenation of the hidden vector of h_{s_i} and $h_{t_{ref}}$, and $v \in \mathbb{R}^{1\times AH}$ is the trainable parameter. B, AH, and C are batch size, attention hidden size and source corpus size respectively. After computing the attention weights, we integrated two hidden vectors (h_{S_1}, h_{S_2}) from S_1 and S_2 to synthesize the target samples (T_{fake}) as:

$$h_{stack} = \begin{bmatrix} h_1 \\ h_2 \end{bmatrix} \quad (13)$$

$$h_{fake} = \alpha_S \cdot h_{stack} \quad (14)$$

$$T_{fake} = Dec\_T(h_{fake}) \quad (15)$$

where h_{stack} is the vertical concatenation of h_1 and h_2.

3.3.3 Emotion Consistency Constraint

To further strengthen the contribution of the synthetic target-aware samples to our main emotion recognition task, we imposed an additional emotion constraint while training a conventional cycleGAN which is inspired by [25] to guarantee that the reconstruction of each source sample preserves the original emotion information and is correctly classified by the well-trained classifier (pre-augmentation) on both source corpora. The constraints are defined as:

$$\mathcal{L}_{Emotion} = \sum_i y_i \log(F_s(G_{T\to S}(G_{S\to T}(S_i)))) \quad (16)$$

where i represents the sample index, y_i is the corresponding annotation for instance S_i from the source corpus, and F_s is the emotion classifier trained by all source datasets.

Therefore, the overall training objective function $\mathcal{L}_{CAEmoCyGAN}$ for our proposed model is defined as,

$$\mathcal{L}_{CAEmoCyGAN} = \lambda_1\mathcal{L}_{cycle} + \lambda_2\mathcal{L}_{identity} + \lambda_3\mathcal{L}_{GAN} + \lambda_4\mathcal{L}_{emotiom} \quad (17)$$

where the λ_1, λ_2, λ_3, and λ_4 are the weights of each specific loss. To further analyze the contribution of each component, we included the CACyGAN in the experiment. The CACyGAN is a special case of the CAEmoCyCAN, in which λ_4 is equal to zero, which means that it does not consider the emotion information reconstruction constraint.

After completing CAEmoCyGAN training, we used it to generate target-aware samples to augment the source data. We then use the augmentation corpus to train an emotion classifier for the target corpus, which was optimized using cross-entropy loss. The overall procedures of inference pseudocode are included in Algorithm 1.

3.4 Target Emotion Classifier

Finally, we aggregated the two sources datasets and the generated synthetic target data as augmented source data (S_aug) to train an emotion classifier (F_{clf}). Using the target-aware synthetic samples, the emotion classifier could more accurately make predictions about the target corpus and increase its variability. In the emotion classifier setting, only the fully connected layers are considered. The common cross-entropy loss was used as our classification criteria, where the classifier loss (L_{clf}) is defined as:

$$\mathcal{L}_{clf} = \sum_i Y_i \log(F_{clf}(S\_aug_i)) \quad (18)$$

where F_{clf} is the final emotion classifier for the target corpus, index i represents the sample index, and S_aug and Y correspond to the augmented source data and their labels, respectively.

Algorithm 1: CAEmoCyGAN Augmentation procedure, m is the batch size

The Augmentation Stage
Data: Random duplication up-sampling of S_1, S_2, and $Target$ to make sure they could be paired.
Result: T_{fake}, Y_{fake}
$n \leftarrow Max(Number\ of\ S_1, Number\ of\ S_2)/m$
$shuffle(S_1, S_2, Target)$
for *batch* = 1, . . . , *n* **do**
  Sample $\{S_1^i\}_{i=1}^m$, a batch from $Source_1$
  Sample $\{S_2^i\}_{i=1}^m$, a batch from $Source_2$
  Sample $\{T^i\}_{i=1}^m$, a batch from $Target$
  $h_1,\ h_2 \leftarrow Enc\_S(S_1),\ Enc\_S(S_2)$
  $h_{ref} \leftarrow Enc\_S(T)$
  $h_{cat} \leftarrow (h_1||h_{ref})||(h_2||h_{ref})$
  $\alpha_S \leftarrow Att(h_{cat})$; ▷ Compute att weights
  $h_{fake} \leftarrow \alpha_S \cdot h_{stack}$; ▷ Integrate hidden
  $T_{fake} \leftarrow Dec\_T(h_{fake})$
  $Y_{fake} \leftarrow \alpha_S \cdot Y_S$
end
The Target Emotion Classifier Training Stage
Data: Initiate an emotion classifier network (F_clf) with parameters θ, and aggregate the original source data with synthetic target data.
Result: F_clf with θ^*
$S\_aug \leftarrow Mix(S_1,\ S_2,\ T_{fake})$
$n \leftarrow S\_aug/m$
for *batch* = 1, . . . , *n* **do**
  Sample $\{S_i\}_{i=1}^m$ a batch from S_{aug}
  Sample $\{Y_i\}_{i=1}^m$ a batch from Y_{aug}
  $\hat{Y} \leftarrow F\_clf(S)$
  $Loss_{clf} \leftarrow CrossEntropy(\hat{Y},\ Y)$
  Update θ
end

4 Experimental Evaluation

In this section, we detail the experimental settings, including evaluation metrics and normalization methods. We conducted the experiment to evaluate two emotion attributes, arousal, and valence, each of which is classified into three classes: low, middle, and high. The following subsections describe the settings and compare the results with those from the baseline models.

4.1 Upper Bound Experimental Setup

According to the setup of our cross-corpus experiments, the upper bound should theoretically correspond to the performance obtained by training and testing all on the target corpus. Here, the MSP-Podcast was used as the target corpus in this work. In the MSP-Podcast, the data are divided into three partitions given by the original corpus: training, validation, and test sets. Using the original settings, we trained a basic DNN feed-forward network with the training set and applied early stopping to the validation set followed by inference on the test set. The two emotion classifier networks were composed of concatenated fully connected layers, whose parameters were set to [1582, 500, 100, 3] with a learning rate of 1e-3 with Adam optimizer and a batch size of 128 with 60 epochs.

4.2 Cross-Corpus Experimental Setup

The input features were 1582 dimensions from the Emobase config in OpenSMILE, and all features were normalized using the min-max normalization scheme. According to section 3, we applied the proposed corpus-aware emotional cycle GAN (CAEmoCyGAN) as our base architecture. The network included two generators: $G_{S\to T}$, which was based on fully connected layers and the ReLU activation functions, and $G_{T\to S}$, in which the encoder mapped the samples from two different source datasets into a hyperplane that represented the source domain and the two distinct decoders mapped the hidden vector back to the two separate source corpora. The detailed parameter settings for the two generators are listed in Table 2. The Adam optimizer and the batch normalization were applied and the batch size was 256 while training all networks.

Due to the differences in the class distribution between the two source and the target dataset, we first conducted random up-sampling to augment each class distribution to become identical. Before training the CAEmoCyGAN, we would pre-trained the network with random pairings of source and target datasets, which meant that each pair were

TABLE 2: CAEmoCyGAN Parameter Settings

Structure	Component	Node Parameters	LR
$G_{S\to T}$	Enc_S	[1582, 1000, 500, 256]	2e-5
	Dec_T	[256, 500, 1000, 1582]	
$G_{T\to S}$	Enc_T	[1582, 1000, 500, 256]	
	Dec_S_1	[256, 500, 1000, 1582]	
	Dec_S_2	[256, 500, 1000, 1582]	
Discriminator	D_S_1	[1582, 1000, 500, 1]	
	D_S_2	[1582, 1000, 500, 1]	
	D_T	[1582, 1000, 500, 1]	
Emotion Classifier	F_{clf}	[1582, 500, 100, 3]	2e-4
Corpus-Aware Att	$hidden\ size$	512	2e-2
	$attn\ size$	128	

organized as [S_1, S_2, T]. Inspired by the cycleGAN, we pre-trained the network over 20 epochs to bi-directionally map between the source and target samples with L1 loss criterion; using this strategy, the network begins training at a stable point rather than using random initial parameters.

To train the main structure, we used three corpus discriminators for *source-IEMOCAP*, *source-VAM* and *target-MSP_Podcast* respectively. The discriminators were composed of a ReLU activation function and multiple fully connected layers. The corpus-aware attention mechanism was also applied, with all the parameters listed in Table 2. To avoid an unstable training process, each component's learning rate was different and corresponded to its characteristics. Notice that the experiment was divided into corpus-aware cycleGAN (CACyGAN) and corpus-aware emotional cycleGAN (CAEmoCyGAN) depending on whether the emotion constraint was included. The weights (λ_1, λ_2, λ_3, λ_4) of cycle loss, identity loss, GAN loss and emotion loss were 10, 10, 1, 10, respectively.

When training the recognition model, we generated 21705 target-aware samples as augmentation data for arousal and 14607 for valence. The different number of augmented samples is due to the aforementioned up-sampling schema applied to the source, e.g., when the task is to recognize arousal, the most prevalent class was middle, with 7235 samples; thus the samples from the low and high classes were randomly up-sampled to 7235 as well, and therefore there were a total of 21705 samples for arousal. We applied the early stopping strategy by validating on a development set to search for a suitable stop point. The stopping metric we used was the UAR performance, which involved randomly sampling 10% of the synthetic target-aware samples from the generative model and using the other 90% for the augmented training set. Notice that in this work, we conduct an unsupervised cross-corpus SER task, which means the classification model would be directly trained on augmented source corpora and then evaluate on the testing partition of the target corpus. Thus, cross-validation schemes are not needed in this setting.

4.3 Baseline Models

Here, we used data augmentation as our main method to mitigate the insufficient data variations in the source datasets. GAN-based generative models have shown a promising capacity to generate samples, and many related algorithms have been proposed recently. In this work, we included a basic cross-corpus DNN model and various generative models which contains cycleGAN, CyCADA, CycleEmotionGAN, and CCEmoGAN as our baseline models for comparison.

- **Cross-corpus DNN**
  This model represents that all the source samples are used to train a recognition model and then test on the target corpus directly. The Cross-corpus DNN model does not include any adaptation scheme which is used as the primitive model for comparison.
- **Random-Paired Auto-encoder (R-P Auto-encoder)**
  In the experiment, we also include the auto-encoder-based model as one of the baseline augmentation models. To make an auto-encoder as a generative model, we randomly paired the source and target sample to form the input pair, e.g., (S_i, T_j). In this way, the auto-encoder model then learns the data distribution between source and target. In the inference stage, we randomly pick source samples to generate fake target samples as a data augmentation method.
- **CycleGAN**
  This method was first proposed in [33], and its primary concept is to learn a non-linear bi-directional mapping function between the source and target corpus. The model has demonstrated successful target synthesis by considering the distribution of the source and target corpus simultaneously. It achieves this goal by unpairing inputs and its unsupervised training procedure. The generator and discriminator parameters are the same as those in our proposed model, to facilitate fair comparison.
- **CyCADA**
  An extension of the traditional cycleGAN, CyCADA [39] includes the main task loss as an optimization criterion for the source corpus. Furthermore, an extra feature discriminator was imposed to distinguish whether the representation was extracted from the source or target. By also simultaneously considering the main task loss, the classifier is effectively trained. After aggregating the information of main task loss and domain-invariant feature representation for both the source and target dataset, the classifier would also show accurate recognition on the target corpus directly.
- **CycleEmotionGAN**
  Similar to the CyCADA, the CycleEmotionGAN [40] strengthens the bi-directional mapping using an emotion constraint. It also considers the emotional semantic consistency loss, i.e., it ensures that the source sample is included in the same class after reconstruction. The proposed model exhibits outstanding performance in the cross-corpus SER task. After training, the generator synthesizes the adapted samples that can be used for data augmentation.
- **CCEmoGAN**
  Instead of searching for a common space to mitigate the domain shift issue, the conditional cycle emotion GAN (CCEmoGAN) [26] was proposed to augment the training data with synthetic fake target samples and constrain the synthetic sample with a conditional one-hot encoded vector, which could effectively compensate for the distribution imbalance in the source

TABLE 3: Proposed Model Comparison on Target MSP-Podcast

	Arousal						Valence					
	IEM2MSP_P		VAM2MSP_P		[IEM+VAM]2MSP_P		IEM2MSP_P		VAM2MSP_P		[IEM+VAM]2MSP_P	
Model Name	UA	WA	UA	WA	UA	WA	UA	WA	UA	WA	UA	WA
Cross-corpus DNN	50.25	46.55	56.92	53.59	53.77	45.41	36.59	22.81	34.83	53.87	42.04	33.82
R-P Auto-encoder	45.89	31.24	58.77	42.32	52.26	35.20	43.28	37.98	38.68	39.74	42.79	38.98
CycleGAN	47.66	40.71	58.51	42.17	52.32	40.09	**43.86**	33.06	39.13	45.84	43.93	35.59
CCEmoGAN	50.42	34.99	**59.98**	41.25	51.23	42.01	43.49	27.85	**39.20**	35.36	43.90	38.38
CyCADA	**51.35**	41.19	57.22	38.91	54.55	38.10	38.82	21.91	34.26	33.79	43.39	41.49
CyEmoGAN	47.33	45.29	57.62	47.43	53.92	46.60	38.20	25.38	35.14	52.45	41.33	35.75
CAEmoCyGAN	-	-	-	-	**61.64**	51.02	-	-	-	-	**44.62**	42.31

TABLE 4: Within Corpus Result of Target Dataset

	Arousal		Valence	
	UA	WA	UA	WA
MSP Podcast	68.85	57.54	44.52	49.55

training samples. CCEmoGAN demonstrates outstanding performances when conducting SER across widely used speech emotion corpora, such as the IEMOCAP, the CIT, and the MSP-IMPROV. Since the data augmentation of Type B (from source to target) fake samples outperforms other models as well as other types of synthetic samples, in this work, we focus on comparing the fake samples that were generated from the source to the target (Type B).

5 Experimental Results and Analysis

In all our experiments, unweighted average recall (UAR) was used as the metric for evaluating the performance of emotion recognition accuracy, and we also present the weighted accuracy (WA) for reference.

5.1 Upper Bound Experiment Result

Here, we run a simple DNN model within corpus following the label division settings in [48], and we obtain similar results. We take this as our upper bound, i.e., training directly on the target corpus. The results are presented in the Table 4. We found that the UARs for the MSP-Podcast were 68.85% for arousal and 44.52% for valence. From Table 4, we observe that the arousal performance overwhelmingly surpassed that of valence, which reinforces that the valence recognition is still a challenging task in this target corpus.

Our proposed model achieved 61.64% and 44.62% for arousal and valence, respectively, on the MSP-Podcast. From these results, we can see that the valence accuracy is almost the same as the upper bound, which means that the result of data augmentation by the proposed CAEmoCyGAN is comparable to the scenario of training within the entire dataset. This demonstrates the effectiveness of our proposed model and may guarantee improvement not only for valence but also for the arousal attribute.

5.2 Baseline Comparisons

All of the results from the baseline models and our proposed model are listed in Table 3. Here, we list three different columns for each emotional attribute: *IEMOCAP-to-TARGET*, *VAM-to-TARGET* and *[IEMOCAP+VAM]-to-TARGET*. In the following paragraphs, we compare the performance in terms of single-source transfer and multi-source transfer.

5.2.1 Single-Source Transfer

For the Cross-corpus DNN results concerning arousal, shown in Table 3, the performance on VAM2MSP_P is 56.92%, which surpasses the result on IEM2MSP_P by 6.67% in absolutely points. In the Cross-corpus DNN model, the source corpus was directly used to train without any adaptation schema, and this result reveals that the VAM dataset was much closer to the target in its original form, and therefore Cross-corpus DNN works better when using the VAM as a source than to the IEMOCAP. We believe it is because of the VAM collection scenarios, which include recordings from TV programs. They are rich in spontaneous dialogue, and therefore more similar to the MSP-Podcast, which includes recordings from online podcasts. Therefore, similar results were obtained by examining other single source augmentation methods, and the best result on VAM2MSP_P was achieved by CCEmoGAN, with an UAR of 59.98%, and the best result on IEM2MSP_P was achieved by CyCADA, with an UAR of 51.35%.

Additionally, we further compared the Cross-corpus DNN result to the best adaptive augmentation methods for both scenarios, and the absolute point improvements were 1.10% for IEM2MSP_P (CyCADA vs. DNN) and 3.06% for VAM2MSP_P (CCEmoGAN vs. DNN). These results show that, when the VAM is the source, the models benefit more from the augmentation methods. We believe the primary reason for this result is the large difference in the amount of data between the IEMOCAP and the VAM (the IEMOCAP has almost tenfold more Table 1). The relatively large amount of data in the IEMOCAP dataset provides needed variability; this is not true of the VAM, even though its affective characteristics seem to be more similar to the target corpus, the MSP-Podcast. This phenomenon also shows that the smaller dataset could benefit much more from augmenting with target-aware samples, as they would further compensate for the lack of diversity and would likely result in higher performance than those datasets with a larger amount of data originally, e.g., the IEMOCAP.

For valence, however, in contrast to the arousal results, both the Cross-corpus DNN model and adaptive augmentation models performed better when using the IEMOCAP as the source dataset. The Cross-corpus DNN's UAR was 36.59% on IEM2MSP_P and 34.83% on VAM2MSP_P. The

TABLE 5: Ablation Study

Model Name	Arousal		Valence	
Target : MSP-Podcast	UA	WA	UA	WA
CycleGAN	52.32	40.09	43.93	35.59
CACyGAN	57.83	45.51	44.33	36.55
CAEmoCyGAN	61.64	51.02	44.62	42.31

TABLE 6: Arousal Recall Result

	CAEmoCyGAN	CACyGAN
Low Recall	**75.38%**	72.96%
Mid Recall	**37.84%**	29.29%
High Recall	71.16%	**71.36%**
UAR	**61.46%**	57.87%
WA	**51.50%**	46.27%

UAR of the best adaptive augmentation model was 43.86% on IEM2MSP_P (CycleGAN) and 39.20% on VAM2MSP_P (CCEmoGAN). The lower accuracy implies that recognizing valence in speech is remains a difficult task. The absolute improvements were 7.27% for IEM2MSP_P and 4.37% for VAM2MSP_P respectively. We hypothesize that this result may be due to the severe label imbalance issue for the valence attribute in the VAM corpus, which only contained 10 samples in the high valence, which accounted for only 1.06% of all data (Table 1). Such imbalanced label distribution could dramatically degrade the performance when conducting SER across corpora. This result also implies that the distribution and amount of data in each class significantly affect the impact of data augmentation for single-source transfer. The above experiments indicate that both the data amount and the data distribution among each class are major considerations when training a model for SER across corpora. The model performance improves only when both of these factors are properly controlled.

5.2.2 Multi-Source Transfer

For the arousal attribute, after aggregating both source datasets to train the Cross-corpus DNN model, the UAR (53.77%) indeed increased when compared with IEM2MSP_P (50.25%), but slightly decreased when compared with VAM2MSP_P (56.92%). Apparently, the larger amount of data in the IEMOCAP dominates the major performance; however, the multi-source training result (Cross-corpus DNN-53.77%) was still better than the single-transfer result from IEM2MSP_P (CyCADA-51.35%) and gained a 2.42% increase in the UAR in terms of absolute points. These improvements imply that the VAM dataset increases the target-relevant emotional speech variability for the IEMOCAP dataset and contributes to the unsupervised recognition model. However, these results are still lower than that of the VAM itself by using (Cross-corpus DNN-56.92%) due to the large proportion of the IEMOCAP samples. When comparing the multi-source results after applying baseline adaptive augmentation models, the best result was 54.55% from CyCADA, which is 0.78% higher than that of the Cross-corpus DNN model, and again shows the superiority of augmentation methods. Our proposed model, CAEmoCyGAN achieved a UAR of 61.64% using multi-source settings and outperformed all the baseline models. When compared with the best baseline multi-source result, CAEmoCyGAN improves by 7.09% when compared to CyCADA. CAEmoCyGAN is also superior to all single-source transfer models.

For the valence, the multi-source training Cross-corpus DNN model showed improved results as arousal, and achieved a 42.04% UAR, which was 5.45% and 7.21% better than that on IEM2MSP_P and VAM2MSP_P, respectively. One particular observation unique to the valence attribute is that the Cross-corpus DNN model's performance was better than both single-source transfer Cross-corpus DNN models on IEM2MSP_P and VAM2MSP_P. A probable reason for this result is that the lack of samples in the VAM indeed hindered the model when making predictions for the more difficult valence recognition task; however, after combining both source samples, the VAM dataset provided the target-relevant valence information and the IEMOCAP guaranteed a sufficient number of training samples. Regarding contributions by both source datasets, the results of using the combined dataset are better than those of the single-source transfer models. Further, under multi-source settings, the traditional CycleGAN augmentation model yields a better accuracy at 43.93% UAR, which is 1.89% higher than that of the Cross-corpus DNN model. Meanwhile, our proposed CAEmoCyGAN still provides the best performance, with 44.62% UAR when comparing with all the baseline models and obtains 2.58% improvement in terms of absolute points.

It is worth mentioning that our proposed CAEmoGAN maintains a high WA while increasing the UAR for both the arousal and valence attributes, which were 51.02% and 42.31%, respectively. This result implies that the proposed augmentation model learns the smaller class size categories better while retaining its high performance for the dominant class. From the above analysis of both emotion attributes, the augmentation method mitigates the mismatch between different sources in both single-source transfer models and multi-source transfer models. Furthermore, our proposed model utilizes the characteristics from both sources to generate the target-aware samples and provides the best results when compared with other augmentation models, which did not perform better than the single-source transfer models.

5.3 Ablation Study

In order to understand the contribution of each component to our proposed model, we divided them into three parts: the traditional cycleGAN, the corpus-aware attention mechanism, and the additional emotion consistency constraint loss. We present the ablation study results in Table 5. When using the MSP-Podcast as the target, the arousal prediction performance increased by 5.51% and 3.81% with the additional corpus-aware attention mechanism and the emotion consistency constraint, respectively. Regarding the more complex emotion, valence, the performance increased by 0.39% and 0.29%, respectively.

Therefore, when learning to integrate characteristics from each source corpus, the synthetic sample augmentation increased the performance across corpora. When considering the emotion consistency constraint, which maintains the emotional information after reconstruction, the performance further improves while conducting emotion recognition.

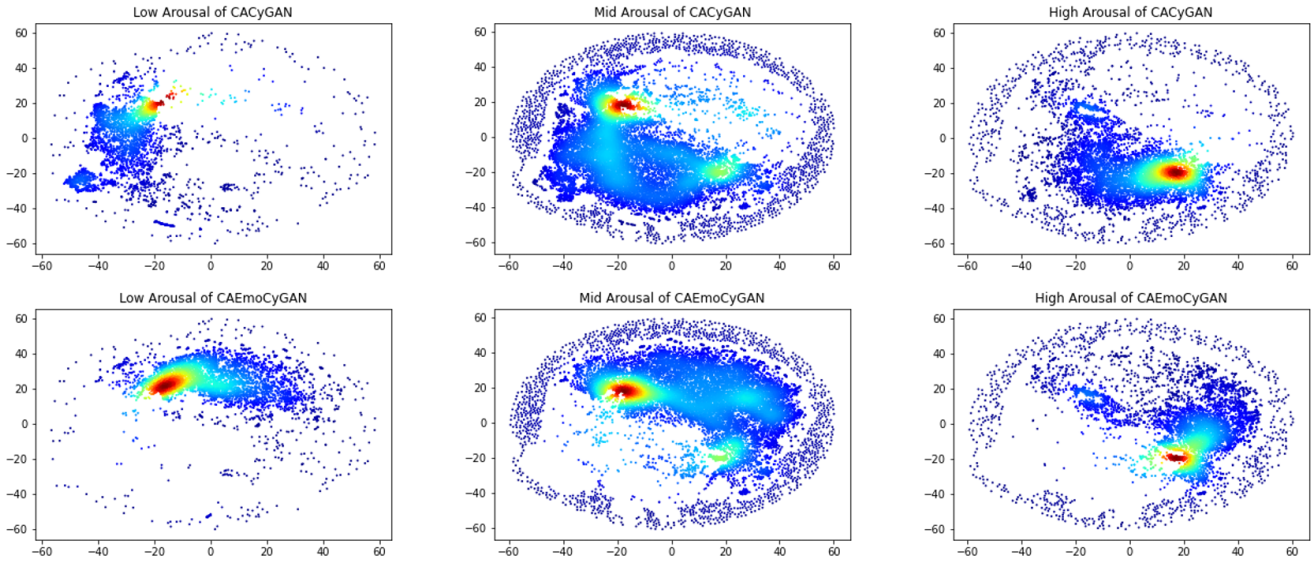


Fig. 2: The t-SNE visualization of the data distribution in arousal when using the MSP-Podcast as the target corpus. The upper row depicts the overlapping density distribution between CACyGAN and the target corpus among three classes, and the lower row is between CAEmoCyGAN and the target corpus as well. The dark blue represents a single point, and the red color represents the high-density area which means a high overlapping region.

5.4 Representation Visualization

In this section, in order to discuss the similarity between different target-aware samples, we visualized the learned representation from CAEmoCyGAN, CACyGAN, and the target to intuitively present the distribution of the target-aware samples and the true target samples. Here, we separated both emotion attributes into three classes, resulting in a total of six classes. The classes can be directly analyzed by examining the overlapping areas in the distribution in Fig. 2.

The representation dimension was reduced using the t-SNE unsupervised algorithm. From Fig. 2, we found that the overlapping part of the distribution between CAEmoCyGAN's synthetic target-aware samples and the true target samples for arousal classes is larger than that between CACyGAN's synthetic target-aware samples and the true target samples in all the classes of arousal. Our model's better ability to synthesize target-aware samples (for every emotion class) further guarantees the robust performance of conducting SER in an unsupervised manner.

We further present the per-class recall rates in TABLE 6. It is evident that the Mid and the Low class of our CAEmoCyGAN improve more compared to CACyGAN, i.e., Low recall improves 2.42% in absolute points (from 72.96% to 75.38%), and Mid recall improves 8.55% in absolute points (from 29.29% to 37.84%). However, only the recall for High slightly drops 0.2% in absolute points. This result indicates the overlapping of low and middle samples between true target and CAEmoCyGAN generated samples are more significant than CACyGAN (see Fig 2).

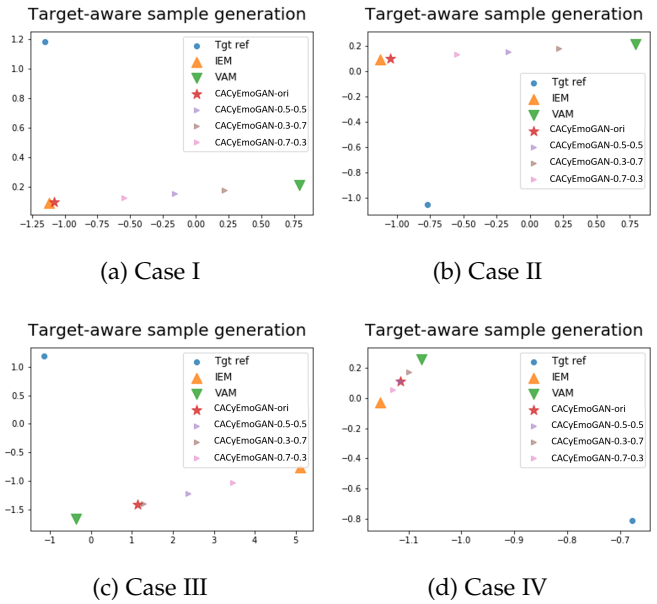
5.5 Learned Attention Weights Visualization

In this section, we further analyze the learned attention weights to investigate the attention learned between the two source datasets and the reference target sample by visualizing the relationship between them. Due to the high number of permutations that would result if we considered all different cases. We first chose four cases sample for visualization to demonstrate the underlying working mechanism of our corpus-aware attention mechanism. These cases are included in Fig. 3, and are described below.

- Case I -
  The label of the reference target sample is equal to one of that from the sources' sample (the IEMOCAP: high, the VAM: low, the MSP-Podcast: high).
- Case II -
  The three samples all come from different classes (the IEMOCAP: high, the VAM: low, the MSP-Podcast: middle).
- Case III -
  The reference MSP-Podcast (high) differs from the sources class, but the sources from the IEMOCAP (low) and the VAM (low) are in the same class.
- Case IV -
  All samples come from the same class (high).

Therefore, for Case I, as seen in Fig. 3a, the reference target MSP-Podcast sample and the IEMOCAP sample belong to the high class, but the VAM sample belongs to the low class. In this case, we found that the target-aware sample integrated using these two source samples is closer to the IEMOCAP than the VAM, and the combination weights among the two samples are 0.98 and 0.02, respectively. This result shows evidence that category alignment is still needed even when the VAM setting is closer to that of the target corpus, as its class is too different from the reference target sample. The category alignment results in attention weights of 0.98 and 0.02 for the IEMOCAP and the VAM, respectively.

Case II extends the above case, meaning that we examine those cases where the emotion category from both sources and the reference target samples to be different. We would



(a) Case I (b) Case II

(c) Case III (d) Case IV

Fig. 3: Corpus-aware attention weights analysis among four different cases. The upward triangle represents the sample from IEMOCAP, the downward triangle represents the VAM sample, the rightward triangle represents different ratio combinations, the blue circle represents the reference target sample, and the red star represents the learned synthetic target-aware sample, which was integrated by the corpus-aware attentions.

TABLE 7: Case-wise Attention Distribution Analysis

	Arousal		Valence	
	IEM>0.5	VAM>0.5	IEM>0.5	VAM>0.5
CASE-I-IEM	**71.33%**	28.67%	**62.65%**	37.35%
CASE-I-VAM	**50.80%**	49.20%	41.25%	**58.75%**
CASE-II	**58.27%**	41.73%	**65.02%**	34.98%
CASE-III	**61.56%**	38.44%	**63.30%**	36.70%
CASE-IV	**64.61%**	35.39%	**59.06%**	40.94%

like to investigate which category from which source is closer to a specific reference target sample class. In Fig. 3b, when the three samples all come from different classes, the reference MSP-Podcast sample is in the middle class; the IEMOCAP sample is in the high class; the VAM sample is in the low class. The attention weights from the source samples are overwhelmingly 0.96 for the IEMOCAP and 0.04 for the VAM, which may imply that model when referencing on the middle class of MSP-Podcast leverages more the IEMOCAP (potentially due to its larger size) than the VAM sample when forming this synthetic case sample.

In Case III as shown in Fig. 3c, when the reference MSP-Podcast (high class) differs from the sources' class, but the sources from the IEMOCAP (low class) and the VAM (low class) are both the same class, the weight distributions are 0.28 for the IEMOCAP and 0.72 for the VAM. Though the reference targets come from different categories, the corpus-aware attention mechanism estimates the needed integration weights. This case may also imply while both sources are under the same category, the corpus collection setting plays a crucial role in directing attention.

For Case IV, when all the samples belong to the same emotion category, each source sample's property is emphasized. In Fig. 3d, all samples from the same category (high class) and the weight for the IEMOCAP is 0.52 while that for the VAM is 0.48. The sample from IEMOCAP gained a little more attention weight and played as a key role during integration and the sample from the VAM also contributed significantly to the synthetic sample. These results indicate to better generate the target-aware samples, fusing different sources is beneficial.

From the visualization in Fig. 3, we find that the corpus-aware attention mechanism seems to learn the attention weights from two sources that are inversely proportional to the distance from the projection of the target reference point to the sources. Considering these four cases, the combination from different emotion categories effectively mitigates the label distortion between the source and target datasets.

5.5.1 *Attention Weights Distribution Analysis*

Following the above analysis, we additionally compute the attention weights distributions (at the corpus level) of synthetic samples among these four cases for both emotion dimensions, and the statistics are shown in Table 7. Here, we separate CASE I into CASE-I-IEM, and CASE-I-VAM which means the label of the reference target sample is the same as the samples from the IEMOCAP or the VAM, respectively. In Table 7, we present quantitative results on how the target samples rely distinctly on both source corpora for each case.

For example, while focusing on valence in CASE-I, we find that if the reference target sample is in the same class of the IEMOCAP, 62.65% of samples would gain more attention weights on the IEMOCAP which means the combination attention weights on the IEMOCAP would be greater than 0.5. On the other hand, if the reference target sample is in the same class of the VAM, 41.25% of the samples would have larger attention weights on the VAM, and the remaining 58.75% would have larger attention weights on the IEMOCAP.

The same tendency is also observed for the arousal dimension, and we find that most of the cases are relying more on the IEMOCAP no matter whether the reference target emotion is in the same emotion class as the source. From the results, we see that the target MSP-Podcast dataset may be more similar to the IEMOCAP than the VAM. The results are also intuitive pleasing, due to the fact that the IEMOCAP is a larger dyadic interaction emotion corpus that contains more useful variability than the VAM. However, the environment of the MSP-Podcast might be more similar to VAM which is recorded from TV programs.

However, it is interesting to observe the results of valence from CASE-I-VAM, where the proportion of higher attention on the VAM (58.75%) surpasses the one in arousal (49.20%). This phenomenon may imply that the valence of the VAM is more similar to the MSP_Podcast than arousal due to the higher weights proportion observed (the higher reliance on the VAM for arousal is 49.20% and the higher reliance on the VAM for valence is 58.75%). This analysis provides preliminary insights on the underlying working mechanism of our corpus-aware attention mechanism and indicates the differences in the source datasets when learning to generate target-aware samples. We also observe a differential effect of our multi-source attention mechanism

TABLE 8: Model Comparison on Target MSP-IMPROV

	Arousal		Valence	
	[IEM+VAM]2MSP_I		[IEM+VAM]2MSP_I	
Model Name	UA	WA	UA	WA
Cross-corpus DNN	63.51	49.51	48.73	56.02
R-P Auto-encoder	57.24	32.95	47.86	46.91
CycleGAN	64.00	45.87	49.74	47.54
CCEmoGAN	62.10	40.06	48.60	40.89
CyCADA	63.83	45.37	44.52	59.65
CyEmoGAN	61.38	52.83	46.10	42.51
CAEmoCyGAN	**65.20**	46.04	**50.06**	46.41

with respect to each emotion dimension (arousal and valence). These insights provide a direction for us to work on in the future.

5.6 Extending Source/Target Experiments

In this section, we conduct an additional experiment to examine the scalability and robustness of our proposed CAEmoCyGAN, by extending the use of our framework with three extra source corpora and one target corpus.

- Additional Target
  The MSP_IMPROV [49] is included as our extra target corpus, which consists of 12 speakers in English and splits into six sessions. All 8438 utterances are labeled with arousal, valence, and dominance as well.
- Additional Source
  Here, we include three common speech emotion corpora to be our extra source corpus, which are the EMODB [50], the MSP_IMPROV, and the CreativeIT [51], respectively. The EMODB is a German emotional dataset, which contains 535 utterances labeled in categorical emotions from 10 speakers (five males, and five females). We map the categorical emotion to three classes following the rule in [15] and only consider neutrality as the middle class. The CreativeIT is an English speech emotional database collected from USC, and all the subjects are encouraged to behave in goal-oriented affective interactions. All 2163 utterances are annotated with arousal and valence with a continuous scale ranging from -1 to 1. These corpora are regularly used in SER tasks.

For the additional target corpus-MSP_IMRPOV, the experiment results are shown in Table 8. From Table 8, our proposed model surpasses other baseline models by 1.2% absolute points in arousal and 0.32% absolute points in valence.

Moreover, in order to extend to the multiple source scenario, besides using the IEMOCAP and the VAM, we include the third source corpus and present the performances obtained in Table 9. When using the MSP_Podcast as the target, from Table 9, we find that the arousal is only slightly bit lower than the two-sources setting, but the results of our proposed model are still better than other baseline models under the two-sources setting. Additionally, more importantly, we find the result of valence recognition, while generally showing a similar tendency as arousal, the result of including the MSP_IMPROV as the third source helps improve the valence recognition to 46.62% (surpassing 44.62% under the two-sources setting).

TABLE 9: Experiment for Extra Source Corpus

	Arousal		Valence	
	[IEM+VAM]2MSP_P		[IEM+VAM]2MSP_P	
Extra Source	UAR	WA	UAR	WA
-	**61.64**	51.02	44.62	42.31
EMODB	56.59	44.34	42.85	43.78
MSP_IMRPOV	54.24	46.98	**46.62**	41.04
CreativeIT	54.22	49.25	42.53	41.18
	[IEM+VAM]2MSP_I		[IEM+VAM]2MSP_I	
Extra Source	UAR	WA	UAR	WA
-	65.20	46.04	50.06	46.41
EMODB	**65.34**	43.01	50.10	48.25
MSP_Podcast	64.84	45.80	**50.67**	51.63
CreativeIT	64.41	54.29	46.91	48.52

When the target dataset is the MSP_IMPROV, the results of arousal are similar to that of the MSP_Podcast. Valence shows better performance by including the EMODB and the MSP_Podcast. The valence accuracy by including the EMODB and the MSP_Podcast are 50.10% and 50.67% in UAR, which are all better than CyCAEmoGAN under the two-sources setting.

From the experiment results, we find that the improvements in the arousal recognition saturate after two sources, but additional valence performance gains are observed by including a third source. These experiments demonstrate the ease of extension in using our proposed CyCAEmoGAN to handle multiple-sources settings, and the inclusion of more source datasets seems to be beneficial for cross-corpus valence recognition.

6 Conclusion

In this work, we propose a novel generative architecture that augments data while conducting SER across corpora in a multiple-source scenario. Data augmentation has been demonstrated to gain advantages from increasing the amount and variability in the training data for ASR. For SER, the most common method employed finds the domain-invariant representation for both the source and target domain, though it might be doable for single source and single target. We observe that the prior state-of-the-art methods cannot perform well when the corpora are unique and diverse, as the IEMOCAP and the VAM are. In order to train the model using all of the unique source datasets we had, we proposed to utilize the characteristics of each unique corpus using a corpus-aware attention mechanism to synthesize the target sample. According to the results, the performance of data augmentation through our proposed model surpassed the most recent state-of-the-art methods. To solve the sparse and contextualized distribution of speech emotion datasets, data augmentation is an imperative method that compensates for the lack of speech emotion corpora as well as the high cost of labeled data acquisition.

Since our proposed generative model aims to consider the uniqueness of each source corpora and learn weights that facilitate corpus integration, the synthetic samples are more likely to mimic the target corpus. Using our corpus-aware attention mechanism, the learned weights fuse the source samples to represent the target sample, which is then more flexible than the one-to-one mapping employed by traditional methods. As our method considers additional emotional characteristics, the performance is further

improved. In this paper, we reached competitive cross-corpus unsupervised emotion recognition performance by integrating information from the two source datasets and analyzing the learned weights of each. In the future, we would like to investigate methods to filter the synthetic samples instead of using all of them and continuously focusing on the more challenging valence recognition task. Further, we will design a more systematic way to integrate more source corpora, which would lead us to investigate the explicit relationship between corpus contexts while acoustic feature similarity is preferred by attention mechanism, i.e., scripted/improvisation, on-site/recording, and even language preferences.

REFERENCES

[1] F. Ren and C. Quan, "Linguistic-based emotion analysis and recognition for measuring consumer satisfaction: an application of affective computing," *Information Technology and Management*, vol. 13, no. 4, pp. 321–332, 2012.

[2] G. N. Yannakakis, "Enhancing health care via affective computing," 2018.

[3] J. Hernandez, R. R. Morris, and R. W. Picard, "Call center stress recognition with person-specific models," in *International Conference on Affective Computing and Intelligent Interaction*. Springer, 2011, pp. 125–134.

[4] A. Menychtas, M. Galliakis, P. Tsanakas, and I. Maglogiannis, "Real-time integration of emotion analysis into homecare platforms," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2019, pp. 3468–3471.

[5] H. Basanta, Y.-P. Huang, and T.-T. Lee, "Assistive design for elderly living ambient using voice and gesture recognition system," in *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2017, pp. 840–845.

[6] E. Polyakov, M. Mazhanov, A. Rolich, L. Voskov, M. Kachalova, and S. Polyakov, "Investigation and development of the intelligent voice assistant for the internet of things using machine learning," in *2018 Moscow Workshop on Electronic and Networking Technologies (MWENT)*. IEEE, 2018, pp. 1–5.

[7] C. Filippini, D. Perpetuini, D. Cardone, A. M. Chiarelli, and A. Merla, "Thermal infrared imaging-based affective computing and its application to facilitate human robot interaction: a review," *Applied Sciences*, vol. 10, no. 8, p. 2924, 2020.

[8] R. K. Moore, "Is spoken language all-or-nothing? implications for future speech-based human-machine interaction," in *Dialogues with Social Robots*. Springer, 2017, pp. 281–291.

[9] C. Tschöpe, F. Duckhorn, M. Huber, W. Meyer, and M. Wolff, "A cognitive user interface for a multi-modal human-machine interaction," in *International Conference on Speech and Computer*. Springer, 2018, pp. 707–717.

[10] P. Song, W. Zheng, S. Ou, X. Zhang, Y. Jin, J. Liu, and Y. Yu, "Cross-corpus speech emotion recognition based on transfer non-negative matrix factorization," *Speech Communication*, vol. 83, pp. 34–41, 2016.

[11] S. Albanie, A. Nagrani, A. Vedaldi, and A. Zisserman, "Emotion recognition in speech using cross-modal transfer in the wild," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 292–301.

[12] Y. Zong, W. Zheng, T. Zhang, and X. Huang, "Cross-corpus speech emotion recognition based on domain-adaptive least-squares regression," *IEEE signal processing letters*, vol. 23, no. 5, pp. 585–589, 2016.

[13] S. Latif, R. Rana, S. Younis, J. Qadir, and J. Epps, "Cross corpus speech emotion classification-an effective transfer learning technique," *arXiv preprint arXiv:1801.06353*, 2018.

[14] Z. Zhang, F. Weninger, M. Wöllmer, and B. Schuller, "Unsupervised learning in cross-corpus acoustic emotion recognition," in *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*. IEEE, 2011, pp. 523–528.

[15] 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.

[16] Z. Huang, W. Xue, Q. Mao, and Y. Zhan, "Unsupervised domain adaptation for speech emotion recognition using pcanet," *Multimedia Tools and Applications*, vol. 76, no. 5, pp. 6785–6799, 2017.

[17] P. Song, "Transfer linear subspace learning for cross-corpus speech emotion recognition." *IEEE Trans. Affect. Comput.*, vol. 10, no. 2, pp. 265–275, 2019.

[18] 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.

[19] M. Abdelwahab and C. Busso, "Domain adversarial for acoustic emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 12, pp. 2423–2435, 2018.

[20] G.-Y. Chao, Y.-S. Lin, C.-M. Chang, and C.-C. Lee, "Enforcing semantic consistency for cross corpus valence regression from speech using adversarial discrepancy learning." in *INTERSPEECH*, 2019, pp. 1681–1685.

[21] P. Wei, Y. Ke, X. Qu, and T.-Y. Leong, "Subdomain adaptation with manifolds discrepancy alignment," *IEEE Transactions on Cybernetics*, 2021.

[22] 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 *Interspeech*, 2019, pp. 171–175.

[23] C. Etienne, G. Fidanza, A. Petrovskii, L. Devillers, and B. Schmauch, "Cnn+ lstm architecture for speech emotion recognition with data augmentation," in *Workshop on Speech, Music and Mind 2018*. ISCA, 2018, pp. 21–25.

[24] L. Yi and M.-W. Mak, "Adversarial data augmentation network for speech emotion recognition," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2019, pp. 529–534.

[25] 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.

[26] B.-H. Su and C.-C. Lee, "A conditional cycle emotion gan for cross corpus speech emotion recognition," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 351–357.

[27] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[28] T.-S. Nguyen, S. Stueker, J. Niehues, and A. Waibel, "Improving sequence-to-sequence speech recognition training with on-the-fly data augmentation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7689–7693.

[29] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5220–5224.

[30] W.-N. Hsu, Y. Zhang, and J. Glass, "Unsupervised domain adaptation for robust speech recognition via variational autoencoder-based data augmentation," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 16–23.

[31] P. Sheng, Z. Yang, and Y. Qian, "Gans for children: A generative data augmentation strategy for children speech recognition," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 129–135.

[32] Z. Chen, A. Rosenberg, Y. Zhang, G. Wang, B. Ramabhadran, and P. J. Moreno, "Improving speech recognition using gan-based speech synthesis and contrastive unspoken text selection," *Proc. Interspeech 2020*, pp. 556–560, 2020.

[33] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[34] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International conference on machine learning*. PMLR, 2017, pp. 214–223.

[35] Y. Xiao, H. Zhao, and T. Li, "Learning class-aligned and generalized domain-invariant representations for speech emotion recognition," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 4, no. 4, pp. 480–489, 2020.

[36] 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.
[37] J. Deng, X. Xu, Z. Zhang, S. Frühholz, and B. Schuller, "Universum autoencoder-based domain adaptation for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 24, no. 4, pp. 500–504, 2017.
[38] M. Neumann and N. T. Vu, "Improving speech emotion recognition with unsupervised representation learning on unlabeled speech," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7390–7394.
[39] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell, "Cycada: Cycle-consistent adversarial domain adaptation," in *International conference on machine learning*. PMLR, 2018, pp. 1989–1998.
[40] S. Zhao, C. Lin, P. Xu, S. Zhao, Y. Guo, R. Krishna, G. Ding, and K. Keutzer, "Cycleemotiongan: Emotional semantic consistency preserved cyclegan for adapting image emotions," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 2620–2627.
[41] 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, pp. 335–359, 2008.
[42] M. Grimm, K. Kroschel, and S. Narayanan, "The vera am mittag german audio-visual emotional speech database," in *2008 IEEE international conference on multimedia and expo*. IEEE, 2008, pp. 865–868.
[43] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, 2017.
[44] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.
[45] F. Haider, S. Pollak, P. Albert, and S. Luz, "Emotion recognition in low-resource settings: An evaluation of automatic feature selection methods," *Computer Speech & Language*, vol. 65, p. 101119, 2021.
[46] S. T. Rajamani, K. T. Rajamani, A. Mallol-Ragolta, S. Liu, and B. Schuller, "A novel attention-based gated recurrent unit and its efficacy in speech emotion recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6294–6298.
[47] C. Fu, C. Liu, C. T. Ishi, and H. Ishiguro, "Maec: Multi-instance learning with an adversarial auto-encoder-based classifier for speech emotion recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6299–6303.
[48] A. Triantafyllopoulos and B. W. Schuller, "The role of task and acoustic similarity in audio transfer learning: Insights from the speech emotion recognition case," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7268–7272.
[49] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. M. Provost, "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.
[50] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, B. Weiss *et al.*, "A database of german emotional speech." in *Interspeech*, vol. 5, 2005, pp. 1517–1520.
[51] A. Metallinou, C.-C. Lee, C. Busso, S. Carnicke, S. Narayanan *et al.*, "The usc creativeit database: A multimodal database of theatrical improvisation," *Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*, p. 55, 2010.

Bo-Hao Su (S'20) is currently pursuing his Ph.D. degree and received his B.S. degree in in the Department of Electrical Engineering at National Tsing Hua University, Taiwan in 2017. He was awarded with NTHU Principal Outstanding Student Scholarship (2018 - 2022), Interspeech 2018 Sub Challenge Championship. His research field includes behavioral signal processing (BSP), cross corpus speech emotion recognition, and machine learning. He is also a student member of ISCA.

Chi-Chun Lee (M'13, SM'20) is an Associate Professor at the Department of Electrical Engineering with joint appointment at the Institute of Communication Engineering of the National Tsing Hua University (NTHU), Taiwan. He received his B.S. and Ph.D. degree both in Electrical Engineering from the University of Southern California, USA in 2007 and 2012. His research interests are in speech and language, affective computing, health analytics, and behavioral signal processing. He is an associate editor for the IEEE Transaction on Affective Computing (2020-), the IEEE Transaction on Multimedia (2019-2020), the Journal of Computer Speech and Language (2021-), and a TPC member for APSIPA IVM and MLDA committee. He serves as an area chair for Interspeech 2016, 2018, 2019, senior program committee for ACII 2017, 2019, publicity chair for ACM ICMI 2018, late breaking result chair for ACM ICMI 2023, sponsorship and special session chair for ISCSLP 2018, 2020.

He is the recipient of the Foundation of Outstanding Scholar's Young Innovator Award (2020), the CIEE Outstanding Young Electrical Engineer Award (2020), the IICM K. T. Li Young Researcher Award (2020), the NTHU Industry Collaboration Excellence Award (2020), and the MOST Futuretek Breakthrough Award (2018, 2019). He led a team to the 1st place in Emotion Challenge in Interspeech 2009, and with his students won the 1st place in Styrian Dialect and Baby Sound subchallenge in Interspeech 2019. He is a coauthor on the best paper award/finalist in Interspeech 2008, Interspeech 2010, IEEE EMBC 2018, Interspeech 2018, IEEE EMBC 2019, APSIPA ASC 2019, IEEE EMBC 2020, and the most cited paper published in 2013 in Journal of Speech Communication. He is a ACM and ISCA member.