Self supervised adversarial domain adaptation for cross-corpus and cross-language speech emotion recognition

Siddique Latif, Rajib Rana, Sara Khalifa, Raja Jurdak, Björn W. Schuller

Self Supervised Adversarial Domain Adaptation for Cross-Corpus and Cross-Language Speech Emotion Recognition

Siddique Latif ⬤, Rajib Rana ⬤, Sara Khalifa ⬤, Raja Jurdak ⬤, *Senior Member, IEEE*, and Björn Schuller ⬤, *Fellow, IEEE*

Abstract—Despite the recent advancement in speech emotion recognition (SER) within a single corpus setting, the performance of these SER systems degrades significantly for cross-corpus and cross-language scenarios. The key reason is the lack of generalisation in SER systems towards unseen conditions, which causes them to perform poorly in cross-corpus and cross-language settings. Recent studies focus on utilising adversarial methods to learn domain generalised representation for improving cross-corpus and cross-language SER to address this issue. However, many of these methods only focus on cross-corpus SER without addressing the cross-language SER performance degradation due to a larger domain gap between source and target language data. This contribution proposes an adversarial dual discriminator (ADDi) network that uses the three-players adversarial game to learn generalised representations without requiring any target data labels. We also introduce a self-supervised ADDi (sADDi) network that utilises self-supervised pre-training with unlabelled data. We propose synthetic data generation as a pretext task in sADDi, enabling the network to produce emotionally discriminative and domain invariant representations and providing complementary synthetic data to augment the system. The proposed model is rigorously evaluated using five publicly available datasets in three languages and compared with multiple studies on cross-corpus and cross-language SER. Experimental results demonstrate that the proposed model achieves improved performance compared to the state-of-the-art methods.

Index Terms—Speech emotion recognition, self-supervised learning, domain adaptation, adversarial learning

1 INTRODUCTION

SPEECH Emotion Recognition (SER) is widely explored by researchers to enable effective human-computer interaction. Speech is a major affect display, and it contains information about emotional expressions that can be automatically identified using machine learning (ML) models. SER systems can help businesses by improving their service delivery. Speech emotion identification can be used in call centres to track customer and agent reactions. Speech-based affect recognition can be effectively utilised in healthcare for diagnosis and monitoring of depression, distress, and bipolar disorder in patients [1],

- *Siddique Latif is with the University of Southern Queensland (USQ), Springfield, QLD 4301, Australia, and also with Distributed Sensing Systems Group, Data61, CSIRO, Pullenvale, QLD 4069, Australia. E-mail: siddique.latif@usq.edu.au.*
- *Rajib Rana is with the University of Southern Queensland (USQ), Springfield, QLD 4301, Australia. E-mail: rajib.rana@usq.edu.au.*
- *Sara Khalifa is with Distributed Sensing Systems Group, Data61, CSIRO, Pullenvale, QLD 4069, Australia. E-mail: sara.khalifa@data61.csiro.au.*
- *Raja Jurdak is with TruNets – Trusted Networks Lab, Queensland University of Technology (QUT), Brisbane, QLD 4000, Australia. E-mail: r.jurdak@qut.edu.au.*
- *Björn Schuller is with GLAM – The Group on Language, Audio, and Music, Imperial College London, SW7 2BX London, U.K., and also with the ZDB Chair of Embedded Intelligence for Health Care and WellbeingUniversity of Augsburg86159 Augsburg, Germany. E-mail: schuller@tum.de.*

(Corresponding author: Siddique Latif.)

[2]. Many other sectors including smart cars [3], forensic sciences [4], education [5], to name a few, are also aiming to utilise SER techniques to improve their performances.

Over the past few years, deep learning (DL) based architectures including deep belief networks (DBN) [6], convolutional neural networks (CNN) [7], and long short term memory (LSTM) networks [8] have significantly improved SER performance compared to the classical machine learning (ML) approaches [9], [10], [11]. SER systems based on deep neural networks (DNNs) perform satisfactorily when training and test data belong to the same corpus. The performance of these systems plummets significantly when the training speech corpus is very different from the testing corpus – known as cross-corpus SER.

One of the key reasons for poor performance in cross-corpus SER is the difference between training and testing speech data distributions. These differences become more prevalent when training and testing data belong to different languages (cross-language SER). As a solution to the problem, researchers use diverse corpora (including multilingual data) for training to create more generalised and robust SER systems [12]. Studies show that an SER model trained on multiple corpora can achieve improved results [9]. However, acoustic training using multiple labelled data is not feasible for all languages, as we have speech corpora in very few languages compared to the number of languages spoken around the world [13], [14] and getting samples for adaptation in rarely spoken languages is challenging.

An alternative and more practical approach to address the above challenge is domain adaptation, which generalises SER systems to unseen conditions by minimising domain shift—the gap between source and target data distributions. Domain adaptation approaches maximise the domain confusion to learn a common feature space by minimising some measures of domain shift such as (a) maximum mean discrepancy [15], [16] or (b) correlation distances [17], [18]. Reconstruction of the target domain using source representation is another way to create a shared representation [19], [20]. These approaches are effectively used in the computer vision domain. However, achieving domain adaptation in speech emotion is more complex, as it requires keeping the emotional information while reducing domain shift in source and target data.

Adversarial domain adaptation methods have become a popular manifestation in SER research to minimise an approximate domain discrepancy distance through an adversarial loss. These methods are closely related to the generative adversarial network (GAN) [21] training, which pits a generator and a discriminator against each other. The generator is trained to generate fake data in a way that confuses the discriminator. In adversarial domain adaptation, this principle is used among the feature encoder, and domain discriminator [22]. In SER, different studies (e.g., [23], [24]) use domain discriminator-based adaptation approaches. However, it is difficult to capture all the useful information and complex structures (such as the emotions) in the feature and label spaces using a *single* domain discriminator [25].

This paper proposes an Adversarial Dual Discriminator (ADDi) network to learn a domain generalised emotional representation that improves cross-corpus and cross-language SER. Our proposed model is equipped with a dual discriminator, which is not explored in SER. This enables the proposed model to generate the domain invariant representations with a three-players adversarial game among generator and dual discriminator.

To address the challenge of limited labelled data, we further propose self-supervised learning for ADDi – we call it sADDi. We propose synthetic data generation as our pretext task – we utilise the unlabelled data to pre-train the encoder component to learn to produce features for emotional synthetic data generation, which can be used to augment the system and help minimise the required labelled training data.

Most of the existing SER studies (e.g., [23], [24], [26]) on adversarial domain adaptation do not consider cross-language, creating a research gap. This is likely due to the complexity of learning a generalised representation for cross-language SER. In this paper, we consider improving cross-language SER.

We summarise the contributions of this paper below. We,

1) propose a novel adversarial domain adaptation technique: ADDi for cross-corpus and cross-language SER. ADDi, for the first time, introduces a dual discriminator for SER, enabling the generation of domain invariant representations with a three-players adversarial game among generator and dual discriminator.

2) enable self-supervised learning with ADDi (we call it sADDi) by generating synthetic data as a pretext task

that effectively utilises unlabelled data to improve the performance and produce synthetic data to augment the SER system.

3) use five widely applied publicly available datasets to comprehensively evaluate the performance of ADDi and sADDi for cross-corpus and cross-language SER performance. Results show that ADDi outperforms the state-of-the-art methods and when including self-supervised learning in ADDi, sADDi offers even higher performance improvement than the state-of-the-art methods. Besides improving the performance, sADDi reduces the required amount of source labelled data by 15-20 % compared to the current and most relevant study (ADDoG) [26] while achieving comparable classification accuracy.

2 RELATED WORK

2.1 Cross-Corpus and Cross-Language SER

Cross-corpus speech emotion recognition is an important task to enable real-life SER applications. It aims to build systems with improved generalisation to perform SER not only in variations in speaker and languages but also in unknown target conditions, including changes in recording environments, noise levels, and elicitation strategy. State-of-the-art SER systems trained on a single corpus fail to perform well in cross-corpus settings. Previous studies explore various techniques to achieve better performance in cross-corpus SER. Schuller *et al.* [12] find that the SER performance degrades due to the acoustic and annotation differences. They perform experiments using six corpora to gain generalisation. They also evaluate multiple normalisation techniques and z-normalisation to achieve the best results. Eyben *et al.* [35] perform cross-corpus SER evaluations using speech databases with realistic and non-prompted emotions. They use a uni-variate ranking of the low-level descriptors (LLDs) to find the most important features and achieve improvement in some settings. They highlight that future efforts are required to address the inconsistencies among multiple corpora by carefully selecting annotations. Zhang *et al.* [27] evaluate unsupervised learning and feature normalisation for cross-corpus SER. They show that adding unlabelled data to agglomerate multi-corpus training sets and utterance level feature normalisation can improve performance. In [36], the authors show the effect of data agglomeration and decision-level fusion for cross-corpus SER. They use six datasets and demonstrate that joint training with multiple corpora and late fusion could help improve performance. These studies show the preliminary feasibility of cross-corpus learning and motivate further in-depth research.

Researchers also explore different techniques to perform emotion identification in cross-language settings. Albornoz *et al.* [37] consider emotion profile-based ensemble support vector machines (SVM) for emotion classification in multiple languages. They model each language independently to preserve the cultural properties and apply the universality of emotions to map and predict emotions in different languages. They use the RML corpus [38] and achieve improved results using their model in a language-independent SER. Li *et al.* [39] develop a three-layered model of acoustic features,

semantic primitives, and emotion dimensions to perform cross-language emotion classification. They apply feature selection and speaker normalisation and evaluate the proposed framework on Japanese, German, Chinese, and English emotional speech corpora. They achieve multilingual recognition performance comparable with a monolingual emotion recogniser. In [14], the authors evaluate cross-lingual SER and highlight the ways of designing an adaptive emotion recognition system for languages with a small available dataset. They show that training the model with multiple languages data can deliver comparable results with a model trained with monolingual data and that augmentation of the training set with a fraction of target language labelled data can help improve the performance. Various other studies (e.g., [35], [36], [40]) explore cross-lingual SER, however, these studies evaluate classical ML models on relatively smaller datasets.

Most recent studies on SER utilise deep representation learning techniques over low-level features. Particularly, studies use deep networks to learn generalised representations to improve performance. For instance, the authors in [9] use DBNs for learning generalised features across multiple datasets. They evaluate the proposed model using six emotional corpora and showed that DBN can provide better performance in cross-corpus SER. They also observe that a DBN can learn a robust representation from many language datasets that helps improve SER performance. In [29], the authors train an attentive convolutional neural network (ACNN) for binary classification of arousal and valence in cross-language and multi-language training settings using French and English language datasets. They show that multilingual training can enhance the performance of the system. Also, they find that the ACNN can be fine-tuned using a fraction of target language data to produce sound results for cross-language SER. Ning *et al.* [41] employ multilingual Bidirectional Long Short-Term Memory (BLSTM) with the shared hidden layers across different languages for universal feature representation learning. They evaluate the proposed model for English and Mandarin corpora and found that cross-lingual knowledge learning using shared hidden layers helps improve SER performance compared to BLSTM variants without shared hidden layers. Some other studies (e.g., [42], [43], [44]) also exploit deep networks to improve cross-corpus and cross-language emotion detection. In general, the methods proposed in these studies require large aggregated speech labelled corpora to achieve generalisation for improved cross-corpus performance. Models training using aggregated corpora is not feasible in real life, as it requires multiple labelled datasets. In contrast, domain adaptation is a more practical approach that improves the system's generalisation without the need for multiple labelled corpora. We review the studies on domain adaptation in the next subsection.

2.2 Adversarial Domain Adaptation

Deep domain adaptation aims to improve the generalisation of SER systems by addressing the problem of domain shift among source and target datasets. Researchers explore different domain adaptation models

(e.g., [45], [46], [47], [48]) to improve cross-corpus and cross-lingual SER. To this end, adversarial domain adaptation techniques are becoming very popular in SER. For instance, the authors in [30] use domain adversarial neural networks (DANN) [22] for cross-corpus emotional attributes' prediction. They learn generalised representations between the source and target data by using a gradient reversal layer (GRL) which propagates back the negative of the gradient produced by the domain classifier to the shared network. They find that the DANN can learn domain invariant representations to cross-corpus SER. Xiao *et al.* [23] propose an adversarial network for class-aligned and generalised domain adaptation. They also exploit GRL to generalise representations among source and target data. They evaluate the proposed model against cross-corpus settings using IEMOCAP and MSP-IMPROV corpora and achieved improved results compared to DANN and AE-based deep architectures. Zhou *et al.* [49] present a class-wise domain adversarial adaptation method to learn common representation to address cross-corpus mismatch issues. They evaluated the proposed model on two datasets including AIBO and EMO-DB for the French language and show that the proposed model achieves better results when training is performed on target data with minimal labels for positive and negative emotion classes recognition. Gideon *et al.*[26] introduce an adversarial discriminative domain generalisation model that follows a "meet in the middle" approach for cross-corpus emotion recognition. The proposed approach utilises the critic network that enables the model to improve the cross-corpus generalisation by iteratively moving representations closer to source and target data. They perform evaluations using English datasets including IEMOCAP, MSP-IMPROV, and PRIORI emotion datasets [26] and show that the proposed framework generates generalised representations for improved cross-corpus SER.

Most of the studies above evaluate adversarial domain adaptation methods for cross-corpus SER using similar language corpora. However, few studies show the effectiveness of their methods for different languages in cross-corpus SER. Ahn *et al.* [34] propose a few shots learning-based unsupervised domain adaptation techniques to learn emotional similarity among source and target domains. They evaluate the proposed model in three different languages and achieve improved results. However, their proposed method requires additional labelled training data to improve the generalisation. Latif *et al.* [31] present a GAN-based adversarial method to learn language invariant representations and evaluate the model for different language datasets. They train support vector machines (SVM) on language invariant representations to improve the performance of cross-language SER. In contrast to these studies, we propose an Adversarial Dual Discriminator (ADDi) network that utilises a dual discriminator to learn generalised representations to improve cross-corpus SER. One of the novel features of our model is the utilisation of self-supervised learning (SSL) for domain adaptation, which has not been explored for SER domain adaptation. Few studies exploited SSL for improving SER performance within-corpus settings. We discuss these studies in the following subsection.

TABLE 1
Summary of a Comparative Analysis of Our Paper With That of the Existing Literature

| Author (Year) | Technique | Evaluations | | Adversarial Learning | Self-Supervised Learning |
		Cross-Corpus	Cross-Language		
Schuller *et al.* (2010) [12]	Feature normalisation	✓	✓	✗	✗
Zhang *et al.* (2011) [27]	Feature normalisation	✓	✓	✗	✗
Kim *et al.* (2017) [28]	Aggregated corpora training	✓	✓	✗	✗
Latif *et al.* (2018) [9]	Transfer learning	✓	✓	✗	✗
Neumann *et al.* (2018) [29]	Aggregated corpora training	✓	✓	✗	✗
Abdelwahab *et al.* (2018) [30]	Domain adaptation	✓	✗	✓	✗
Latif *et al.* [31] (2019)	Domain adaptation	✗	✓	✓	✗
Song *et al.* [32] (2019)	Feature subspace learning	✓	✓	✗	✗
Gideon *et al.* (2019) [26]	Domain adaptation	✓	✗	✓	✗
Xiao *et al.* (2020) [23]	Domain adaptation	✓	✗	✓	✗
Luo *et al.* [33] (2019)	Feature subspace learning	✓	✓	✗	✗
Ahn *et al.* (2021) [34]	Domain adaptation & aggregated corpora training	✓	✓	✓	✗
Ours (2022)	Domain adaptation	✓	✓	✓	✓

Very few studies have used Adversarial Learning for cross-language SER. (checked green).

2.3 Self-Supervised SER

Self-supervised learning (SSL) [50] is a new paradigm in ML, which uses data for supervision. The self-supervised task, also known as the pretext task, uses the unlabelled data to guide downstream tasks. SSL-based models are getting tremendous interest in computer vision [51], natural language processing (NLP) [52], and automatic speech recognition (SER) [53]; however, few studies utilise SSL in SER. In [54], the authors propose a multitask SSL technique to learn a shared speech representation, where a single encoder network is followed by multiple workers that jointly solve different self-supervised tasks. They perform evaluations on speaker, phoneme, and emotional cue recognition, and achieve improved results. Self-supervised multimodal representation learning though transformers [55] is increasingly gaining momentum to improve SER [56]. Khare *et al.* [57] use transformer-based SSL to improve the performance of multimodal emotion recognition. They fine-tune a transformer trained on a masked language modelling task and can improve emotion recognition performance by 3 % on the CMU-MOSE dataset [58]. A recent study [59] presents a visually-guided SSL framework for improving the SER performance. The authors generate video frames using still images by conditioning the network on corresponding audio. In this way, the pre-trained encoder part of their network learns important features to generate realistic facial and lip movements. They hypothesise that the features learnt by the encoder are highly correlated with the presence of emotion and particular phonemes. They utilise these representations for ASR and SER to achieve state-of-the-art results. In contrast, we propose to generate synthetic emotional data as a pretext task, which adversarially enables the encoder to encode discriminative features for emotional data generations. We use this encoder for our downstream domain adaptation task, which helps produce emotionally discriminative features while minimising the gap between source and target domains. In addition, the synthetic emotional data generated in our downstream task acts as a by-product that can be utilised to augment the system.

2.4 The Research Gap (Summary)

The related work can be summarised as follows.

- Several studies show that DL models trained using multiple sources corpora can improve cross-corpus SER performance; however, acoustic training from multiple language data in real-life is not a feasible approach due to the unavailability of sufficient labelled data for multiple languages. Therefore, there is a need for new methods to overcome this limitation.
- Adversarial neural networks based domain adaptation approaches are widely used for cross-corpus SER; however, there is still room for performance improvement, particularly for cross-language SER.
- Self-supervised learning can be used as an effective tool to address the limited label issue but has not been fully explored and used for cross-corpus SER.

In Table 1, we contrast our work with the literature, briefly showcasing how we aim to address the research gaps.

3 PROPOSED MODEL

The core of the proposed model is the Adversarial Dual Discriminator (ADDi) network and the module for the pretext task enabling self-supervision. We propose the generation of synthetic data as a pretext task, wherein we essentially pre-train an encoder that we later use to realise the proposed Self-supervised Adversarial Dual Discriminator (sADDi) network.

3.1 Adversarial Dual Discriminator (ADDi) Network

Our proposed Adversarial Dual Discriminator (ADDi) network is equipped with an encoder (E), a generator (G_p), and dual discriminators: (D^s) and (D^t). An overview of the proposed framework is shown in Fig. 1, where subfigure with label 2 shows ADDi network. It performs adversarial domain adaptation by learning domain invariant features. We represent the source domain data and target domain data as $X^s = \{x_i^s, y_i^s\}_{i=1}^n$ and $X^t = \{x_j^t\}_{j=1}^m$, respectively. The encoder (E)
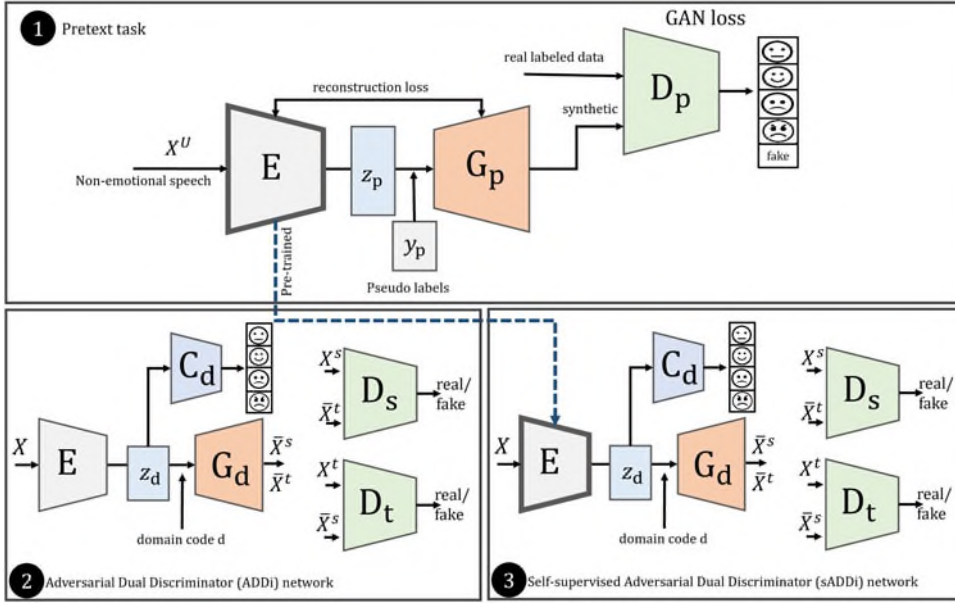
Fig. 1. Overall structure of the proposed framework. We use numbers for different components, where (1) shows the pretext task that is trained on synthetic data generation; (2) represents the Adversarial Dual Discriminator (ADDi) for domain adaptation; and (3) shows the self-supervised Adversarial Dual Discriminator (sADDi), where we use pre-trained encoder from the pretext task as highlighted with a dashed line.

attempts to map the input data either from source or target to a domain invariant latent representation ($z_d = E(X), X \in X^s \cup X^t$). The generator ($G_d$) conditioned on domain code d uses this domain invariant latent representation (z_d) to generate source $\bar{X}^s$ or target $\bar{X}^t$ domain samples. The generator G_d is adversarially connected with two discriminators. The objective function of the generator is as follows:

$$\mathcal{L}_{\mathcal{G}} = \mathcal{L}_{\text{AE}} + \lambda \mathcal{L}_G^{\text{adv}}, \tag{1}$$

$$\mathcal{L}_G^{\text{adv}} = \mathbb{E}\left[\log\left(1 - D_s(G_d(E(X), d))\right)\right] + \mathbb{E}\left[\log\left(1 - D_t(G_d(E(X), d))\right)\right], \tag{2}$$

$$\mathcal{L}_{\mathcal{AE}}(X, G_d(E(X))) = \|X - \bar{X}\|_2^2, \tag{3}$$

where λ is a balancing parameter. The generator is connected to the dual discriminators D_s and D_t, which play the three-players minimax adversarial game to produce z_d to be domain invariant. The generator also acts as the decoder and it reconstructs back the input samples X with the latent representation z_d using the reconstruction loss in Equation (3).

The dual discriminators are tasked to distinguish the real data from the fake data. Particularly, for domain code $d = 0$, the discriminator D^s differentiates between $\tilde{X}^t = G_p(E(X), d)$ (fake) and source data X^s (real), whereas the discriminator D^t discriminates between $\tilde{X}^s = G_p(E(X), d)$ (fake) and source data X^s (real), for domain code $d = 1$. The adversarial process of the generator G_p and two discriminators minimises the divergence between the source and target data distributions and forces the encoder E to generate a generalised latent representation z_d across the source and target domains. The objective function for the dual discriminators can be given as follows:

$$\mathcal{L}_{D_s} = \mathbb{E}\left[-\log D_s(X^s)\right] + \mathbb{E}\left[-\log(1 - D_s(G_d(X, d)))\right]_{\{d=0\}}, \tag{4}$$

$$\mathcal{L}_{D_t} = \mathbb{E}\left[-\log D_t(X^t)\right] + \mathbb{E}\left[-\log(1 - D_t(G_d(X, d)))\right]_{\{d=1\}}. \tag{5}$$

The classifier C_d is connected with the latent representation z_d and minimises the cross entropy loss for emotion

classification during training using only the source data labels and the error is back-propagated through the network to update E. In this way, the encoder E gets influenced by the classifier and enforces z_d to be emotionally discriminating as representation. This helps produce emotionally discriminative and domain invariant representations to perform cross-corpus and cross-language robust SER. When the pre-trained encoder E is fine-tuned in the domain adaptation task, it promotes the discriminative power of a domain invariant representation and further boosts the performance of the system. We discusses the encoder pre-training in the next section.

During training, first, the autoencoder is updated using the Equation (3). Afterwards, G_d is updated to generate the fake samples using z_d and the domain code d. We concatenate the one-hot domain code with the encoded representation z_d and feed to G_d. We further update the discriminators based on the domain codes. For the samples with $d = 0$, D^s is updated, whereas D^t is updated for the samples with $d = 1$. Finally, we update the C_d for the source data samples.

3.2 Pretext: Synthetic Emotional Data Generation and Self Supervised Adversarial Dual Discriminator (sADDi)

A pretext task is used in self-supervised learning (SSL) to generate useful representations that can provide a supervisory signal to the down stream task. It is a predefined task for the network to solve learning the objective function [51]. Most of the SSL pretext tasks are designed based on intuition or heuristics [60]. There is no guarantee on the compatibility between the pretext task and the down stream task [61]. For SER, solving multiple audio based self-supervised tasks can offer improvements [54]. However, these tasks have been evaluated for within corpus SER settings. The design of an SSL pretext task for domain adaption is challenging, as emotionally discriminative generalised representations are required to effectively perform cross-corpus SER.

We use synthetic emotional data generation as pretext task for cross-corpus domain adaptation. The intuition here is that the encoder network pre-trained to encode discriminative features for emotional synthetic data generation when utilised in domain adaptation should help produce an emotionally discriminative generalised representation.

The architecture for our pretext task is shown in Fig. 1 as a subfigure with label 1. It follows the GAN architecture consisting of a generator and a discriminator. Both these networks play an adversarial game defined by the following optimisation program in Equation (6).

$$\min_{G} \max_{D} \mathbb{E}_{x \sim P_{\text{data}}}[\log D(x)] + \mathbb{E}_{x \sim P_G}[\log(1 - D(G(z)))]. \quad (6)$$

The generator network captures the data distribution and generates new samples by incorporating feedback from the discriminator network. The discriminator network in a GAN is simply a classifier. It tries to classify the real and fake data, generated by the generator network. While there are many variants of GAN architectures (e.g., [62], [63], [64]), we use the balancing GAN [65] like architecture due to its effectiveness in SER [66]. It consists of an encoder (E), generator/decoder (G_p), and discriminator (D_p). The encoder network (E) takes the non-emotional speech data (X^U) and generates latent code z_p. We concatenate the z_p with the pseudo labels (y_p) and feed to the generator (G_p) to generate the synthetic data. Since the unlabelled samples do not belong to any emotional class, the pseudo labels in four classes are randomly generated and uniformly distributed to the unlabelled non-emotional speech. In this way, the G_p network conditioned on the (y_p) has explicit emotion class label information during generation like the conditional GAN [67]. During the adversarial training, G_p is tasked to generate samples in different classes based on y_p; and D_p is trained to differentiate the generated samples (by generator (G_p)) as fake and real samples to their class labels. The generator tries to avoid the fake label and matches the desired emotional class labels. The discriminator is optimised to output $N_c + 1$ neurons, where N_c represents the emotional classes (happy, sad, neutral, or angry) and the last neuron represents the fake class as used in [65], [68]. Since the encoder (E) is coupled with the GAN, it learns to encode features for different emotional classes in the latent space of the generator G_p. After pre-training, we fine-tune the encoder network in our ADDi network which we name self-supervised Adversarial Adversarial Dual Discriminator (sADDi). It is highlighted in the Fig. 1 with blue dashed line.

4 EXPERIMENTAL SETUP

4.1 Datasets

To evaluate the performance of our proposed model, we use five different emotional datasets, including IEMOCAP, MSP-IMPROV, RECOLA, EMODB, and FAU-AIBO, which are commonly used for cross-corpus and cross-language emotion classification research [13]. In order to use additional unlabelled data for self-supervised learning (SSL), we use a subset of Librispeech [69], which is a corpus of read English speech, suitable for training and evaluating models on automatic speech and speaker recognition systems. Below, we briefly describe these datasets.

TABLE 2
Mapping Rules for IEMOCAP and RECOLA

Corpus	Low/Negative	High/Positive
IEMOCAP	[1, 2.5]	(2.5, 5]
RECOLA	[-1, 0]	(0, 1]

Here bracket includes the elements listed and the parenthesis does not contain the listed elements.

4.1.1 IEMOCAP

This database contains 12 hours of audiovisual data, including audio, video, textual transcriptions, and facial motion information [70]. The recordings are collected from 10 professional actors (five males and five females) during dyadic interactions. In contrast to reading text with prototypical emotions, dyadic interactions allowed the actors to perform more spontaneous emotion [71]. For categorical labels, each sentence is annotated by three annotators and the participant. Finally, an utterance is assigned a label if at least three annotators are assigned the same label. Overall, IEMOCAP contains nine emotions: excited, happy, sad, neutral, angry, disgust, frustrated, fearful, and surprised. Similarly to previous studies [72], we only use utterances of four categorical emotions, including happy, neutral, sad, and angry in this study by merging "happy" and "excited" as one emotion class "happy". The final dataset includes $5\,531$ utterances ($1\,636$ happy, $1\,708$ neutral, $1\,084$ sad, and $1\,103$ angry instances). For continuous labels, IEMOCAP is also annotated for arousal and valence on a scale of 1 to 5. We map continuous labels to binary labels as presented in Table 2.

4.1.2 MSP-IMPROV

The MSP-IMPROV dataset is an acted audiovisual emotional database recorded from 12 speakers performing dyadic interactions [73]. Overall, the recordings are grouped into six sessions and each session contains the recordings of one male, and one female actor similar to IEMOCAP [70]. The scenarios were carefully designed to control emotion and lexical content while maintaining naturalness in the recordings. The MSP-IMPROV is annotated through perceptual evaluations using crowdsourcing [74]. This corpus contains utterances in four categorical emotions: angry, happy, neutral, and sad. To be consistent with previous studies [10], [75], we use all utterances with four emotions: anger (792), happy (2 ,644), sad (885), and neutral (3 ,477).

4.1.3 RECOLA

RECOLA [76] is a French multimodal corpus of spontaneous collaborative and affective interactions. While solving a collaborative task, speakers recorded the dyadic conversations during a video conference. 46 participants (27 females, and 19 males) were recruited to record this corpus. We use the publicly available portion of RECOLA, which contains 1,308 utterances of 23 speakers. An open-source web-based tool ANNEMO[1] was developed for its affective annotation. RECOLA is annotated with continuous labels, including arousal and valence in the range $[-1, 1]$. We use RECOLA for cross-

---

1. https://diuf.unifr.ch/main/diva/recola/annemo

corpus language SER and perform binary classification of arousal (low/high) and valence (negative/positive) as considered in [29]. Table 2 shows the mapping of original annotations to a binary scheme for IEMOCAP and RECOLA.

4.1.4 EMODB

EMODB [77] is the most popular and widely used publicly available emotional dataset in German Language, recorded by the Institute of Communication Science, Technical University Berlin. It contains audio recordings of seven emotions recorded by ten professional speakers in 10 German sentences. This study selects four basic emotions: happy, sad, neutral, and angry to perform categorical cross-language emotion recognition.

4.1.5 FAU-AIBO

FAU-AIBO [78] corpus is a spontaneous emotional corpus in the German language. It contains 9.2 hours of speech from 51 children from different schools while interacting with Sony's pet robot AIBO. In this study, we select FAU-AIBO to evaluate the proposed framework against completely naturalist emotional speech. We map this corpus to binary valence for evaluations.

4.1.6 LibriSpeech

The LibriSpeech dataset [69] contains $1\,000$ hours of English read speech from $2\,484$ speakers. This corpus is derived from audiobooks and is commonly used for automatic speech and speaker recognition problems [79], [80]. The training portion of LibriSpeech is divided into three subsets, with an approximate recording time of 100, 360 and 500 hours. This paper uses the subset that contains 100 hours of recordings. These recordings span over 251 speakers.

4.2 Speech Features Extraction

We represent the speech sample in Mel Filter Banks (MFBs), a widely used speech representation in speech research [26], [81]. We use the Kaldi speech recognition toolkit [82] to extract 40-dimensional MFBs from each utterance. To extract MFBs, we use default options, including a Povey window with a frame length of 25 ms and a frameshift of 10 ms, a preemphasis coefficient of 0.97, and a low cutoff of 20 Hz. These configurations are selected based on [26] to make a fair comparison. Due to the varying lengths of the audio samples, we pad the MFBs with zeros to the length of the longest emotional utterance.

4.3 Model Configuration

This subsection presents the configuration of three models, including a baseline Convolutional Neural Network (CNN), the proposed Adversarial Dual Discriminator (ADDi) network, and the pretext task GAN. Each of these models takes MFBs as the input feature set. Each experiment considers labelled source data for training, and target data is used for testing. We train all these models using Adam as the optimiser with default parameters and a starting learning rate of 0.0001. We compute the validation accuracy at the end of each epoch during training. If the validation accuracy did not improve after 5 epochs, we restore the model to the best epoch and halves the learning rate. This process continues until the learning rate reaches below 0.00001.

The CNN baseline network comprises a feature encoder and emotion classifier. The feature encoder consists of convolutional and max-pooling layers, whereas the classifier part utilises the fully connected layers for classification. Due to the unavailability of target data in the experiments, it is difficult to validate all the hyperparameters of the network for cross-dataset SER. Therefore, we select the parameters commonly used in prior studies [26], [83], [84]. The feature encoder has three convolutional layers, each followed by the pooling layers. We start with a large filter size of 16 in the first convolutional layer as suggested by prior work [84]. The convolutional layers capture the salient regions within the MFBs and create the feature maps. The pooling layers reduce the dimension of these feature maps by identifying the most relevant features. We use the max-pooling layer to give better performance than average pooling during experiments. The feature encoder encodes the entire utterances into the 256 features. The classifier uses these features for emotion classification. We have two dense layers with hidden units of 256 and 128. We employ a dropout layer between two dense layers with a dropout rate of 0.3 to avoid overfitting.

The ADDi network also has an encoder component to encode input MFBs to the domain invariant representation that is used by the generator. We apply a similar encoder architecture to the baseline CNN. The decoder/generator has three transposed convolutional layers to generate samples using the encoded latent representation. Two discriminators and the classifier of ADDi have two hidden layers containing hidden units of 256 and 128 in number. Like the baseline CNN and ADDi, our pre-training GAN also has an encoder network that follows a similar architecture. We employ the same architecture for the discriminator as for the ADDi network. We select the Rectified Linear Unit (ReLU) as a non-linear activation function for all the models due to its better performance than hyperbolic tangent and leaky ReLU during validation.

5 EXPERIMENTS AND EVALUATIONS

We apply a two-tiered evaluation approach: We evaluate ADDi to understand the significance of the proposed dual-discriminator based framework. We then evaluate sADDi to understand the relative significance of self-supervised learning for ADDi. We evaluate the performance of the proposed ADDi and sADDi networks in cross-corpus and cross-language settings by comparing them with related studies that report similar results. To further extend the extent of our comparison, we implement related models including a CNN (baseline), a GAN [31], a DANN [30], a DBN [9], a CNN-LSTM [43], and an autoencoder-based model as used in [85] and compare our results with these models. We repeat each experiment ten time and calculated mean and standard deviation. Results are presented using the unweighted average recall rate (UAR), a widely accepted metric in the field.

TABLE 3
Cross-Corpus SER Results in UAR (%) Using IEMOCAP (English) and MSP-IMPROV (English)

Model	IEMOCAP to MSP-IMPROV	MSP-IMPROV to IEMOCAP
CNN (baseline)	42.5±1.6	44.3±1.5
DANN [30]	42.8±1.4	44.9±1.7
GAN [31]	43.6±1.3	45.8±1.5
ADDOG [26]	44.4±0.9	47.4±0.7
ADDi (proposed)	**45.1±0.8**	**48.2±0.6**
CNN$_{SSL}$ (baseline)	43.8±1.2	45.3 ± 1.1
sADDi (proposed)	**47.1±0.5**	**49.8±0.6**

5.1 Cross-Corpus Results

We evaluate the proposed ADDi and sADDi networks for cross-corpus SER using the IEMOCAP and MSP-IMPROV datasets. Both of these datasets are recorded in similar laboratory conditions in English. In this experiment, we consider no labels for the target dataset. We use a random 80:20 (train:test) split of the source data and train the model as used in [26]. We compare the performance of the ADDi network with a baseline CNN, ADDOG [26], a DANN and a GAN [31]. The results are presented in Table 3.

Compared to these existing methods and baseline, ADDi achieves better results. ADDi achieves 2.6% and 3.9% relative improvements compared to the baseline CNN for IEMOCAP to MSP-IMPROV and MSP-IMPROV to IEMOCAP experiments, respectively. Amongst the previous studies, ADDOG utilises the critic component similar to a Wasserstein GAN [63] to learn generalised representations for cross-corpus SER, while another study [31] applies a single discriminator based adversarial method to minimise the domain gap, and whereas in [30], a gradient reversal layer (GRL) [22] is used to minimise the gap between the source and target domains. In contrast to these studies, ADDi utilises a dual discriminator based network to learn a domain invariant representation by bringing source and target features closer to each other with three-players adversarial minimax games hence producing better results. Using the ablation study in

Section 5.8, we further quantify the relative significance of our dual discriminator based approach.

Table 3 also shows the self-supervised learning (SSL) for ADDi which we called sADDi above. When we pre-train the encoder component in the sADDi network using the proposed synthetic data generation pretext task, it learns to encode discriminative representation for synthetic emotional data generation through the process of accomplishing the proposed pretext task. This helps to produce emotionally discriminative domain generalised features while fine-tuning the encoder in sADDi and the baseline CNN. Results by the SSL methods are separated with a bold line in Table 3, which shows that the pre-training of the encoder considerably improves the cross-corpus SER. It is worth noting that the performance of the baseline CNN is also improved by utilising the pre-trained encoder by our proposed pretext task, which attests the effectiveness of the proposed self-supervised pretext task.

5.2 Cross-Language Results

We evaluate ADDi and sADDi for cross-language SER using both dimensional and categorical emotions. We use the IEMOCAP and RECOLA datasets for dimensional emotion and perform binary arousal and valence classification. All data from one language is used as a training set and all samples of the respective target language are used as the test set. This is the same evaluation strategy used in [29]. We also implement domain adaptive models, including the GAN and DANN for comparison on IEMOCAP and REC-OLA. Cross-language SER results using IEMOCAP and RECOLA are presented in Table 4.

Using our proposed ADDi framework, we achieve better results compared to [29], where the authors use an Attentive Convolutional Neural Network (ACNN) to achieve promising results by fine-tuning the model on the target language. We also compare our results with domain adaptation architectures, including GAN and DANN in Table 4. Compared to these models, ADDi is able to capture an emotion discriminative generalised representation by adversarially minimising the domain shift among source and target language data to

TABLE 4
Dimensional Cross-Language SER Results by UAR (%) Using the IEMOCAP and RECOLA Datasets

Model	IEMOCAP (English) to RECOLA (French)		RECOLA (French) to IEMOCAP (English)	
	arousal	valence	arousal	valence
CNN (baseline)	59.2± 1.8	48.5± 1.5	60.7± 1.6	48.3± 2.0
ACNN [29]	59.3	49.1	61.2	47.5
GAN [31]	59.8±1.9	49.8±1.7	60.3±1.3	48.7± 1.5
DANN [30]	60.1±2.1	50.2±1.5	61.5±1.5	49.2± 1.4
ADDi (proposed)	**61.5±1.2**	**51.8±1.4**	**62.2±1.3**	**50.9±1.2**
CNN$_{SSL}$ (baseline)	60.1±1.5	49.2± 1.3	61.2 ± 1.6	49.0± 1.4
sADDi (proposed)	**63.8±1.0**	**53.8±1.2**	**64.2±1.4**	**52.5±1.3**
using fraction of target date for fine-tuning.				
ACNN [29] (500 target samples)	67.03	50.42	63.07	49.81
sADDi (250 target samples)	**70.3±1.3**	**57.1±1.3**	**68.6±1.0**	**56.3±1.1**
using language information.				
sADDi (250 target samples)	**72.4±1.6**	**60.1±1.4**	**70.2±1.3**	**59.3±1.2**

TABLE 5
Categorical Cross-Language SER Results by UAR (%) Using
the IEMOCAP and EMODB Datasets

Model	IEMOCAP (English) to EMODB (German)	EMODB (German) to IEMOCAP (English)
CNN (baseline)	42.2± 1.9	38.4±2.2
DBN [9]	42.5±2.1	39.5±2.4
CNN-LSTM [43]	42.1±1.8	38.9±2.1
AE [85]	43.2±2.3	40.1±1.8
GAN [31]	44.3±1.7	40.3±1.7
DANN [30]	43.5±1.8	40.5±2.0
ADDi (proposed)	**46.1±1.6**	**41.2±1.8**
CNN$_{SSL}$ (baseline)	43.5±1.7	40.2±1.9
sADDi (proposed)	**48.3±1.5**	**44.8±1.6**

improve SER across different language data. Performance is further improved when features learnt through SSL are utilised to guide the cross-language domain adaptation using sADDi. It is important to note that the performance of all the models is close to the chance level UAR (i.e., 50 %), which shows the complexity of cross-language SER. However, our model improves the baseline results compared to the previous studies. To further improve the baseline performance, we perform two experiments. In the first experiment, we utilise a fraction of target data in the training set. Results in Table 4 show that including only 250 target language data yields considerable improvements compared to ACNN [29] with 500 target samples. We incorporate language id with the source language data and 250 target language samples in the second experiment. Results are reported in Table 4, which shows that the performance is improved for both arousal and valance prediction using the language information in the training data. However, the language information helps valence prediction more than the arousal prediction, which indicates that valence is more lexically dependent than arousal.

We also compare our results on categorical cross-language emotion classification with different studies [9], [30], [31], [43], [85] and present the results in Table 5. In [9], the authors use transfer learning to improve cross-language SER using DBNs. A CNN-LSTM is suggested in [43] and an autoencoder is tested in [85] for cross-language SER. We also compare our results with GAN and DANN-based domain adaptive implementations for cross-language SER. ADDi achieves better results compared to all, which attests that ADDi learns greater domain generalised representation for cross-language scenarios. Compared to baseline, ADDi achieves 3.9 % and 2.8 % relative improvements for IEMOCAP to EMODB and EMODB to IEMOCAP experiments, respectively. Similar to the dimensional emotions, performance on categorical cross-language SER is further boosted by using a self-supervised ADDi (sADDi) network, which shows that the network is able to produce better generalised features for cross-language SER guided by the synthetic emotional data generation pre-training.

5.3 Impact of Pretext Selection: Reconstruction versus Synthetic Data Generation

We propose synthetic data generation as a pretext task for self-supervised learning. In this experiment, we evaluate the effectiveness of this pretext task by comparing it with reconstruction as a pretext task. We make the comparison for both, the baseline and the ADDi networks.

Reconstruction is widely used as a pretext task in the computer vision literature [51], [86], wherein an autoencoder network is used to reconstruct the input back from the compressed representation. To use reconstruction as a pretext, we use unlabelled data (LibriSpeech) for unsupervised reconstruction and pre-train the encoder component to be utilised in the downstream task. We use the LibriSpeech data to generate the synthetic emotional data and pre-train the encoder component to use synthetic data generation for pretext.

Results of the comparisons are presented in Fig. 2 for cross-corpus and cross-language SER. For cross-corpus SER, we use IEMOCAP and MSO-IMPROV, and IEMOCAP and RECOLA are used for cross-language SER.

Compared to the reconstruction-based pretext task, we achieve better results using synthetic emotional data generation for both the baseline CNN model and ADDi. However, despite the popularity of autoencoder-based reconstruction pretext tasks in computer vision, it could not produce strong representations for transfer tasks in SER. One possible reason might be that the autoencoder only learns to encode abstract bottleneck representations from non-emotional speech, which cannot provide supervisory signals in the downstream domain adaptation task.

5.4 Impact of Data Augmentation

We generate the synthetic data during our pre-training step and use it to augment the source training data. We evaluate the model in a cross-corpus setting using IEMOCAP as training and MSP-IMPROV as testing data to compare the results with [83], [87], [88]. In [87], the authors investigate a GAN to generate the synthetic feature vectors using low dimensional features to augment the SER. Bao et al. [88] apply a CycleGAN based model for synthetic samples by transferring feature vectors extracted from a large unlabelled speech data into the target synthetic emotional samples. They augment the SER system with synthetic features to improve SER performance. Recently, Latif et al. [83] utilise the combination of a GAN and mixup [89] to generate synthetic samples for SER augmentation. Similar to these studies, we also augment the SER system with synthetic data and perform evaluations using real, synthetic, and real plus synthetic data. We also use MSP-IMPROV as the target data, as per these studies. We randomly select 30 % of the data as a development set for hyper-parameter selection and the remaining 70 % as testing data as used in these studies. We keep these splits speaker independent. Results are compared with these studies in Table 6. As expected, the synthetic data alone cannot offer better results, but we get better performance when we augment source data with the synthetic data to train ADDi.

5.5 Impact of Incorporation of Source/Target Data

This experiment incorporates the labelled target data into the training and validation. Here, we present the results using IEMOCAP and MSP-IMPROV in Fig. 3. Similar results are achieved for cross-language datasets. We plot the results with different percentages of target data using the baseline approach and ADDi. ADDi improves the results considerably against baseline CNN in all the case, even for a small

(a) IEMOCAP to MSP-IMPROV

(b) MSP-IMPROVt to IEMOCAP

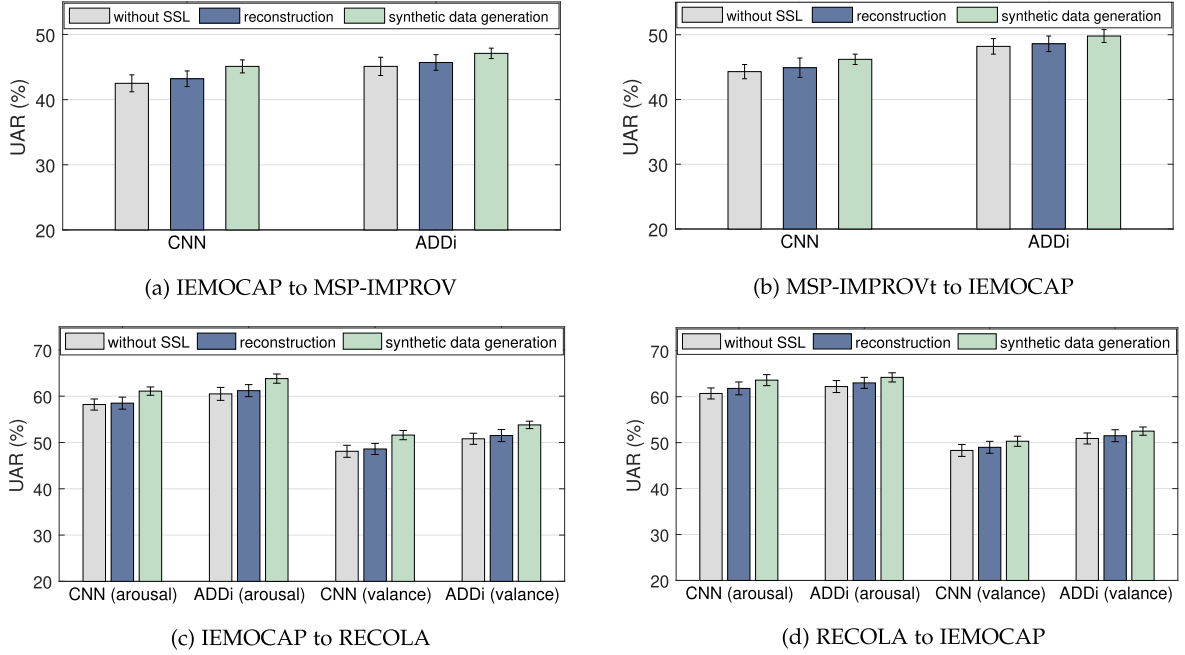(c) IEMOCAP to RECOLA

(d) RECOLA to IEMOCAP

Fig. 2. Impact of self-supervised pre-training on cross-corpus SER (Figs. 2a and 2b) using the IEMOCAP and MSP-IMPROV datasets and cross-language SER (Figs. 2c and 2d) using IEMOCAP and RECOLA.

TABLE 6
UAR Comparison for Cross-Corpus Evaluation Using Synthetic Data to Augment the Model

Studies	Real	Syn.	Real+Syn
Sahu et al. [87]	45.14	33.96	45.40
Bao et al. [88]	45.58 ± 0.40	41.58 ± 1.29	46.5 ± 0.43
Latif et al. [83]	46.0 ± 0.57	42.15 ± 1.12	46.60 ± 0.45
ADDi	**47.83 ± 0.45**	**42.25 ± 0.95**	**48.61 ± 0.40**

We also explore the effect of decreasing the percentage of source data on the performance of cross-corpus SER using IEMOCAP and MSP-IMPROV. In both cases, sADDi performs better than the baseline. We also compare the results with ADDOG [26] in Figs. 3c and 3d. The red dot shows the performance achieved by ADDOG using 100 per cent of source data. We achieve these results using 80-86 % of source data as highlighted by a dotted blue line.

5.6 Evaluations in The Wild

In this section, we evaluate the performance of the proposed model on the naturalist speech. For this experiment, we use FAU-AIBO corpus that contains the natural speech of children in the German language. We perform binary valence

percentage target labelled data. Fig. 3 shows that the margin of UAR improvement decreases with incorporating larger percentages of labelled target data. This may indicate that the generalisation effect diminishes once there is sufficient amount of labelled target domain data.
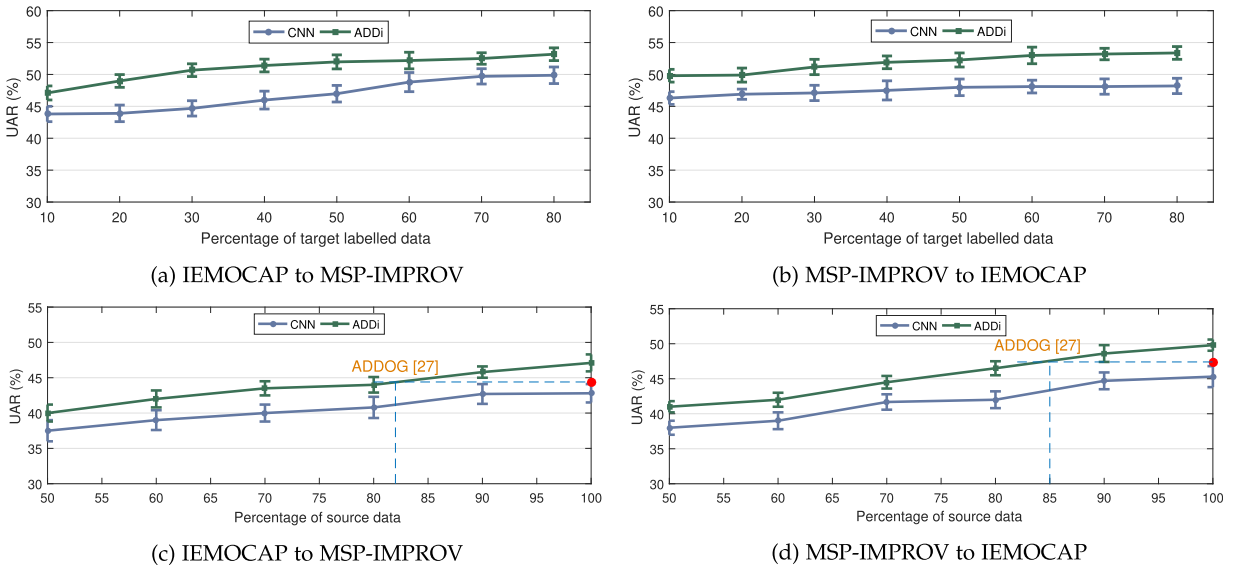

(a) IEMOCAP to MSP-IMPROV

(b) MSP-IMPROV to IEMOCAP

(c) IEMOCAP to MSP-IMPROV

(d) MSP-IMPROV to IEMOCAP

Fig. 3. Results for cross-corpus SER with increasing amounts of labels from the target and source datasets.

TABLE 7
Results With the Naturalist Speech Using FAU-AIBO Corpus

Model	IEMOCAP (English) to FAU-AIBO (German)	FAU-AIBO (German) to IEMOCAP (English)
CNN (baseline)	53.7±1.8	52.1±1.5
DBN [9]	54.5 ±2.0	50.7 ±1.7
sADDi	**58.3 ± 1.5**	**56.9 ±1.6**

classification as used by previous studies [9], [90]. This experiment is more difficult compared to the previous experiments, due to the difference in language, age, and elicitation strategy. We train the model on source data and 20% target data is used as validation and the remaining is used for testing. Results for both experiments are reported in Table 7, which shows that the proposed model considerably improved performance compared to the DBNs [9] and baseline CNNs. sADDi is improving the results by above 4% for both experiments presented in Table 7. This confirms the effectiveness of our sAADi network that can produce generalised representations for evaluations against naturalist speech.

5.7 Size of Pretext Task Training Data

We next examine the effect of the size of training data in pretext tasks on the performance of sADDi. We plot the results for IEMOCAP and MSP-IMPROV in Fig. 4. For both
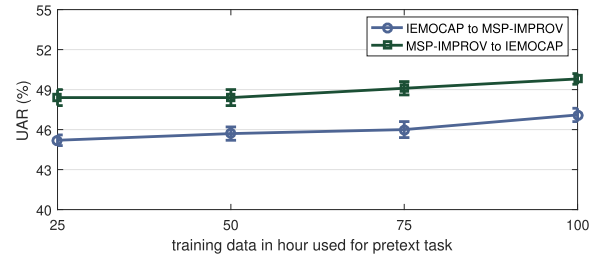

Fig. 4. Effect of increasing the training data in hours on the performance (UAR %).

experiments, we find that the increase of training data in the pretext task helps improve the performance of the downstream emotion classification. This shows that increasing the training data in the pretext task enables the model to produce representations suitable for the downstream task of speech emotion recognition (SER).

5.8 Ablation Experiments

In this experiment, we validate the necessity and effectiveness of each module integrated with our proposed model ADDi. Results are presented in Table 8 for cross-corpus and cross-language evaluations using IEMOCAP to MSP-IMPROV and IEMOCAP to EMODB. These results are computed without any data augmentation and pre-training. Results for synthetic data augmentation and pre-training are presented in Sections 5.4

TABLE 8
Results for Cross-Corpus and Cross-Language Using IEMOCAP to MSP-IMPROV, and IEMOCAP to EMODB, Respectively

Model	Configuration	Discriminators	Decoder	Encoder	Classifier	Cross-corpus UAR (%)	Cross-Language UAR (%)
1		2	✓	✓	✓	**45.1±0.8**	**46.1±1.6**
2		1	✓	✓	✓	43.2±1.3	44.0±1.9
3		1	✓	✓	✓	43.1±1.4	43.5±1.8
4		✗	✓	✓	✓	42.7±1.2	43.3±1.8
5		✗	✗	✓	✓	42.5±1.6	41.2±1.9

and 5.3. This experiment starts with the ADDi model (model 1) that contains all the components, including the encoder, two discriminators, the generator/decoder, and the classifier. We remove one discriminator in models 2 and 3. This makes the model similar to the standard GAN with an additional classifier and autoencoder. We keep removing different modules until we obtain the baseline CNN network (model 5), containing only the encoder and classifier. We also plot the configurations of these models 1-5 in Table 8. There is a considerable drop in UAR when one or more components are removed. When a single discriminator – either D_s or D_t – is used in model 2 or 3, we see a performance drop for both cross-corpus and cross-language SER. This shows that the single discriminator networks cannot achieve better generalisation compared to the three-players adversarial learning performed by the dual discriminator and generator approach in the ADDi network. Similarly, ADDi also achieves considerably improved results compared to models 4 and 5 in cross-corpus and cross-language settings. In models 4 and 5, there is no component (i. e., discriminator) to promote generalised representations in the network by minimising the domain gap between source and target data. This shows that implanting a domain adaptation component in the pipeline of the deep model is important to learn to improve generalised features for cross-corpus and cross-language SER. Overall, these ablation experiments show that all the components in the proposed ADDi models are chosen carefully for effective domain adaptation for SER.

6 CONCLUSION AND FUTURE WORK

This contribution addressed the open challenge of improving the speech emotion recognition (SER) performance in cross-corpus and cross-language settings. We proposed the Adversarial Dual Discriminator (ADDi) network that minimises the domain shift among emotional corpora adversarially. We focused on exploiting the unlabelled data with self-supervised pre-training and proposed self-supervised ADDi (sADDi). For sADDi, we suggested synthetic data generation as a pretext task, which (1) helped improve the domain generalisation performance of an SER system to tackle the larger domain shift between training and test distributions in cross-corpus and cross-language SER; and (2) produced byproduct synthetic emotional data to augment the SER system and minimise the requirement of source labelled data. The key highlights are as follows:

- The introduced dual discriminator based ADDi network offers improved cross-corpus and cross-language SER without using any target data labels compared to the single discriminator and other state-of-the-art approaches. This is mainly due to the dual discriminator using a three-players adversarial game to learn generalised representations.
- Considerable improvements in results were found when partial target labels were fed to the network training. This helped the ADDi to regulate the generalised representations based on the target data by maximally matching the data distributions.
- Our proposed self-supervised pretext task produces synthetic data as a byproduct to augment the system to achieve better performance. We were able to reduce 15-20 % source training data using sADDi

while achieving similar performance reported by a recent related study [26].

Future studies will include evaluating the ADDi and sADDi architectures to model other factors of speech variations, including age, subject, gender, phoneme, noise, and recording device. Further experiments may include evaluating the proposed methods in wild conditions like noisy speech and adversarial noise. We are also interested in exploring multimodal pretext task techniques in our future work. Multimodal human interaction in video and textual form can provide various opportunities for self-supervised learning to improve cross-corpus and cross-language SER.

REFERENCES

[1] S. Latif, J. Qadir, A. Qayyum, M. Usama, and S. Younis, "Speech technology for healthcare: Opportunities, challenges, and state of the art," *IEEE Rev. Biomed. Eng.*, vol. 14, pp. 342–356, 2020.

[2] R. Rana *et al.*, "Automated screening for distress: A perspective for the future," *Eur. J. Cancer Care*, vol. 28, 2019, Art. no. e13033.

[3] H. Detjen, S. Faltaous, B. Pfleging, S. Geisler, and S. Schneegass, "How to increase automated vehicles' acceptance through in-vehicle interaction design: A review," *Int. J. Hum.–Comput. Interaction*, vol. 37, no. 4, pp. 308–330, 2021.

[4] R. Leshem, M. Icht, R. Bentaur, and B. M. Ben-David, "Processing of emotions in speech in forensic patients with schizophrenia: Impairments in identification, selective attention, and integration of speech channels," *Front. Psychiatry*, vol. 11, 2020, Art. no. 1227.

[5] E. Yadegaridehkordi, N. F. B. M. Noor, M. N. B. Ayub, H. B. Affal, and N. B. Hussin, "Affective computing in education: A systematic review and future research," *Comput. Educ.*, vol. 142, 2019, Art. no. 103649.

[6] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.

[7] Y. LeCun *et al.*, "Handwritten digit recognition with a back-propagation network," in *Proc. Adv. Neural Inf. Process. Syst.*, 1990, pp. 396–404.

[8] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[9] S. Latif, R. Rana, S. Younis, J. Qadir, and J. Epps, "Transfer learning for improving speech emotion classification accuracy," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2018, pp. 257–261.

[10] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and J. Epps, "Direct modelling of speech emotion from raw speech," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 3920–3924.

[11] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE Trans. Multimedia*, vol. 16, no. 8, pp. 2203–2213, Dec. 2014.

[12] B. Schuller *et al.*, "Cross-corpus acoustic emotion recognition: Variances and strategies," *IEEE Trans. Affect. Comput.*, vol. 1, no. 2, pp. 119–131, Jul.–Dec. 2010.

[13] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Qadir, and B. W. Schuller, "Survey of deep representation learning for speech emotion recognition," *IEEE Trans. Affect. Comput.*, to be published, doi: 10.1109/TAFFC.2021.3114365.

[14] S. Latif, A. Qayyum, M. Usman, and J. Qadir, "Cross lingual speech emotion recognition: Urdu vs. western languages," in *Proc. Int. Conf. Front. Inf. Technol.*, 2018, pp. 88–93.

[15] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, pp. 97–105.

[16] H. Yan, H. Ding, P. Li, Q. Wang, Y. Xu, and W. Zuo, "Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2272–2281.

[17] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," in *Proc. 30th AAAI Conf. Artif. Intell.*, 2016, pp. 2058–2065.

[18] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 443–450.

[19] M. Ghifary, W. B. Kleijn, M. Zhang, D. Balduzzi, and W. Li, "Deep reconstruction-classification networks for unsupervised domain adaptation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 597–613.

[20] M. Ghifary, W. B. Kleijn, M. Zhang, and D. Balduzzi, "Domain generalization for object recognition with multi-task autoencoders," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2551–2559.

[21] I. Goodfellow et al., "Generative adversarial nets," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[22] Y. Ganin et al., "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2096–2030, 2016.

[23] Y. Xiao, H. Zhao, and T. Li, "Learning class-aligned and generalized domain-invariant representations for speech emotion recognition," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 4, no. 4, pp. 480–489, Aug. 2020.

[24] Y. Yin, B. Huang, Y. Wu, and M. Soleymani, "Speaker-invariant adversarial domain adaptation for emotion recognition," in *Proc. Int. Conf. Multimodal Interaction*, 2020, pp. 481–490.

[25] T. Jing and Z. Ding, "Adversarial dual distinct classifiers for unsupervised domain adaptation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2021, pp. 605–614.

[26] J. Gideon, M. McInnis, and E. M. Provost, "Improving cross-corpus speech emotion recognition with adversarial discriminative domain generalization (ADDoG)," *IEEE Trans. Affect. Comput.*, vol. 12, no. 4, pp. 1055–1068, Oct.–Dec. 2021.

[27] Z. Zhang, F. Weninger, M. Wöllmer, and B. Schuller, "Unsupervised learning in cross-corpus acoustic emotion recognition," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2011, pp. 523–528.

[28] J. Kim, G. Englebienne, K. P. Truong, and V. Evers, "Towards speech emotion recognition "in the wild" using aggregated corpora and deep multi-task learning," in *Proc. 18th Annu. Conf. Int. Speech Commun. Assoc.*, 2017, pp. 1113–1117.

[29] M. Neumann and N. G. Thang Vu , "Cross-lingual and multilingual speech emotion recognition on English and French," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5769–5773.

[30] M. Abdelwahab and C. Busso, "Domain adversarial for acoustic emotion recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 12, pp. 2423–2435, Dec. 2018.

[31] S. Latif, J. Qadir, and M. Bilal, "Unsupervised adversarial domain adaptation for cross-lingual speech emotion recognition," in *Proc. 8th Int. Conf. Affect. Comput. Intell. Interaction*, 2019, pp. 732–737.

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

[33] H. Luo and J. Han, "Nonnegative matrix factorization based transfer subspace learning for cross-corpus speech emotion recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2047–2060, 2020.

[34] Y. Ahn, S. J. Lee, and J. W. Shin, "Cross-corpus speech emotion recognition based on few-shot learning and domain adaptation," *IEEE Signal Process. Lett.*, vol. 28, pp. 1190–1194, 2021.

[35] F. Eyben, A. Batliner, B. Schuller, D. Seppi, and S. Steidl, "Cross-corpus classification of realistic emotions–some pilot experiments," in *Proc. 7th Int. Conf. Lang. Resour. Eval.*, 2010, pp. 77–82.

[36] B. Schuller, Z. Zhang, F. Weninger, and G. Rigoll, "Using multiple databases for training in emotion recognition: To unite or to vote?," in *Proc. 12th Annu. Conf. Int. Speech Commun. Assoc.*, 2011, pp. 1553–1556.

[37] E. M. Albornoz and D. H. Milone, "Emotion recognition in never-seen languages using a novel ensemble method with emotion profiles," *IEEE Trans. Affect. Comput.*, vol. 8, no. 1, pp. 43–53, Jan.–Mar. 2017.

[38] Y. Wang and L. Guan, "Recognizing human emotional state from audiovisual signals," *IEEE Trans. Multimedia*, vol. 10, no. 5, pp. 936–946, Aug. 2008.

[39] X. Li and M. Akagi, "Improving multilingual speech emotion recognition by combining acoustic features in a three-layer model," *Speech Commun.*, vol. 110, pp. 1–12, 2019.

[40] S. M. Feraru, D. Schuller, and B. Schuller, "Cross-language acoustic emotion recognition: An overview and some tendencies," in *Proc. Int. Conf. Affect. Comput. Intell. Interaction*, 2015, pp. 125–131.

[41] Y. Ning et al., "Learning cross-lingual knowledge with multilingual BLSTM for emphasis detection with limited training data," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 5615–5619.

[42] Y. Zhang, Y. Liu, F. Weninger, and B. Schuller, "Multi-task deep neural network with shared hidden layers: Breaking down the wall between emotion representations," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 4990–4994.

[43] J. Parry et al., "Analysis of deep learning architectures for cross-corpus speech emotion recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 1656–1660.

[44] J. Kim, G. Englebienne, K. P. Truong, and V. Evers, "Towards speech emotion recognition "in the wild" using aggregated corpora and deep multi-task learning," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2017, pp. 1113–1117.

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

[46] J. Deng, X. Xu, Z. Zhang, S. Frühholz, and B. Schuller, "Universum autoencoder-based domain adaptation for speech emotion recognition," *IEEE Signal Process. Lett.*, vol. 24, no. 4, pp. 500–504, Apr. 2017.

[47] M. Abdelwahab and C. Busso, "Supervised domain adaptation for emotion recognition from speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 5058–5062.

[48] E. N. N. Ocquaye, Q. Mao, H. Song, G. Xu, and Y. Xue, "Dual exclusive attentive transfer for unsupervised deep convolutional domain adaptation in speech emotion recognition," *IEEE Access*, vol. 7, pp. 93 847–93 857, 2019.

[49] H. Zhou and K. Chen, "Transferable positive/negative speech emotion recognition via class-wise adversarial domain adaptation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 3732–3736.

[50] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, "Self-taught learning: Transfer learning from unlabeled data," in *Proc. 24th Int. Conf. Mach. Learn.*, 2007, pp. 759–766.

[51] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 11, pp. 4037–4058, Nov. 2021.

[52] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A Lite BERT for self-supervised learning of language representations," in *Proc. Int. Conf. Learn. Representations*, 2019.

[53] A. Baevski, S. Schneider, and M. Auli, "VQ-WAV2VEC: Self-supervised learning of discrete speech representations," in *Proc. Int. Conf. Learn. Representations*, 2019.

[54] S. Pascual, M. Ravanelli, J. Serrà, A. Bonafonte, and Y. Bengio, "Learning problem-agnostic speech representations from multiple self-supervised tasks," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 161–165.

[55] A. Vaswani et al., "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[56] M. Macary, M. Tahon, Y. Estève, and A. Rousseau, "On the use of self-supervised pre-trained acoustic and linguistic features for continuous speech emotion recognition," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2021, pp. 373–380.

[57] A. Khare, S. Parthasarathy, and S. Sundaram, "Self-supervised learning with cross-modal transformers for emotion recognition," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2021, pp. 381–388.

[58] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2236–2246.

[59] A. Shukla, K. Vougioukas, P. Ma, S. Petridis, and M. Pantic, "Visually guided self supervised learning of speech representations," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 6299–6303.

[60] L. Xiao et al., "Self-supervised domain adaptation with consistency training," in *Proc. 25th Int. Conf. Pattern Recognit.*, 2021, pp. 6874–6880.

[61] R. Zhang, P. Isola, and A. A. Efros, "Split-brain autoencoders: Unsupervised learning by cross-channel prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1058–1067.

[62] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8789–8797.

[63] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 214–223.

[64] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," in *Proc. Int. Conf. Learn. Representations*, 2018.

[65] G. Mariani, F. Scheidegger, R. Istrate, C. Bekas, and C. Malossi, "BAGAN: Data augmentation with balancing GAN," 2018, *arXiv:1803.09655*.

[66] S. E. Eskimez, D. Dimitriadis, R. Gmyr, and K. Kumanati, "GAN-based data generation for speech emotion recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 3446–3450.

[67] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*.

[68] A. Chatziagapi *et al.*, "Data augmentation using GANs for speech emotion recognition," pp. 171–175, 2019.

[69] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 5206–5210.

[70] C. Busso *et al.*, "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, 2008, Art. no. 335.

[71] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Trans. Affect. Comput.*, vol. 10, no. 4, pp. 471–483, Fourth Quarter 2019.

[72] S. Latif, R. Rana, J. Qadir, and J. Epps, "Variational autoencoders for learning latent representations of speech emotion: A preliminary study," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2018, pp. 3107–3111.

[73] 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 Trans. Affect. Comput.*, vol. 8, no. 1, pp. 67–80, First Quarter 2017.

[74] A. Burmania, S. Parthasarathy, and C. Busso, "Increasing the reliability of crowdsourcing evaluations using online quality assessment," *IEEE Trans. Affect. Comput.*, vol. 7, no. 4, pp. 374–388, Oct.–Dec. 2016.

[75] J. Gideon, S. Khorram, Z. Aldeneh, D. Dimitriadis, and E. M. Provost, "Progressive neural networks for transfer learning in emotion recognition," 2017, *arXiv:1706.03256*.

[76] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit.*, 2013, pp. 1–8.

[77] F. Burkhardt *et al.*, "A database of german emotional speech," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2005, pp. 1517–1520.

[78] S. Steidl, *Automatic Classification of Emotion Related User States in Spontaneous Children's Speech*. Logos-Verlag Berlin, Germany, 2009.

[79] A. Bérard, L. Besacier, A. C. Kocabiyikoglu, and O. Pietquin, "End-to-end automatic speech translation of audiobooks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 6224–6228.

[80] H. Dubey, A. Sangwan, and J. H. Hansen, "Transfer learning using raw waveform sincnet for robust speaker diarization," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6296–6300.

[81] Y. Xu, Q. Kong, Q. Huang, W. Wang, and M. D. Plumbley, "Convolutional gated recurrent neural network incorporating spatial features for audio tagging," in *Proc. Int. Joint Conf. Neural Netw.*, 2017, pp. 3461–3466.

[82] D. Povey *et al.*, "The kaldi speech recognition toolkit," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2011.

[83] S. Latif, M. Asim, R. Rana, S. Khalifa, R. Jurdak, and B. W. Schuller, "Augmenting generative adversarial networks for speech emotion recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 521–525.

[84] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Epps, and B. W. Schuller, "Multi-task semi-supervised adversarial autoencoding for speech emotion recognition," *IEEE Trans. Affect. Comput.*, to be published, doi: 10.1109/TAFFC.2020.2983669.

[85] J. Deng, Z. Zhang, E. Marchi, and B. Schuller, "Sparse autoencoder-based feature transfer learning for speech emotion recognition," in *Proc. Humaine Assoc. Conf. Affect. Comput. Intell. Interaction*, 2013, pp. 511–516.

[86] A. Shukla, S. Petridis, and M. Pantic, "Visual self-supervision by facial reconstruction for speech representation learning," in *Proc. Sight Sound Workshop*, 2020.

[87] S. Sahu, R. Gupta, and C. Espy-Wilson, "On enhancing speech emotion recognition using generative adversarial networks," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2018, pp. 3693–3697.

[88] F. Bao, M. Neumann, and N. T. Vu, "Cyclegan-based emotion style transfer as data augmentation for speech emotion recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 2828–2832.

[89] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *Proc. Int. Conf. Learn. Representations*, 2018.

[90] F. Eyben *et al.*, "The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 190–202, Apr.–Jun. 2016.

Siddique Latif received the bachelor's degree in electronic engineering from International Islamic University, Islamabad, Pakistan, in 2014, sponsored by the National ICT Scholarship Program, the MSc degree in electrical engineering from the National University of Sciences and Technology (NUST), Islamabad, Pakistan, in 2018, and the PhD degree from the University of Southern Queensland (USQ), Australia, and Distributed Sensing Systems Research Group, Data61–CSIRO. Before that, he was working as a research associate with the IHSAN Lab, Information Technology University, Punjab, Lahore, Pakistan. His research work focuses on representation learning, using unlabelled, weakly-labelled, partially-labelled multi-modal data.

Rajib Rana received the BSc degree in computer science and engineering from Khulna University, Bangladesh, with the Prime Minister and President's Gold medal for outstanding achievements and the PhD degree in computer science and engineering from the University of New South Wales, Sydney, Australia, in 2011. He is an experimental computer scientist, Advance Queensland research fellow and a senior lecturer with the University of Southern Queensland, Australia. He is also the director of the IoT Health research program with the University of Southern Queensland, Australia. He is recipient of the prestigious Young Tall Poppy QLD Award 2018 as one of Queensland's most outstanding scientists for achievements in the area of scientific research and communication. His research work aims to capitalise on advancements in technology along with sophisticated information and data processing to better understand disease progression in chronic health conditions and develop predictive algorithms for chronic diseases, such as mental illness and cancer. His current research focus is on unsupervised representation learning. He received his postdoctoral training with Autonomous Systems Laboratory, CSIRO before joining the University of Southern Queensland, Australia, in 2015.

Sara Khalifa received the PhD degree in computer science and engineering from UNSW, Sydney, Australia. She is currently a senior research scientist with the Distributed Sensing Systems research group, Data61-CSIRO. She is also an honourary adjunct lecturer with the University of Queensland, Australia, and conjoint lecturer with the University of New South Wales, Australia. Her research interests rotate around the broad aspects of mobile and ubiquitous computing, mobile sensing, and Internet of Things (IoT). Her PhD dissertation received the 2017 John Makepeace Bennett Award which is awarded by CORE (Computing Research and Education Association of Australasia) to the best PhD dissertation of the year within Australia and New Zealand in the field of computer science. Her research has been recognised by multiple iAwards including 2017 NSW Mobility Innovation of the year, 2017 NSW R&D Innovation of the year, National Merit R&D Innovation of the year, and the Merit R&D award at the Asia Pacific ICT Alliance (APICTA) awards, commonly known as the 'Oscar' of the ICT industry in the Asia Pacific, among others.

Raja Jurdak (Senior Member, IEEE) received the BE degree in computer and communications engineering from the American University of Beirut, Lebanon, in 2000, the MS degree in computer networks and distributed computing from the Electrical and Computer Engineering Department, University of California, Irvine, Irvine, California, in 2001, and the PhD degree in information and computer science, University of California, Irvine, Irvine, California, in 2005. He is a professor of distributed systems and chair in applied data sciences with the Queensland University of Technology, Australia, and director of the Trusted Networks Lab. He previously established and led the Distributed Sensing Systems Group with CSIRO's Data61. He also spent time as visiting academic with MIT and Oxford University, in 2011 and 2017. His research interests include trust, mobility and energy-efficiency in networks. He has published more than 200 peer-reviewed publications, including two authored books most recently on blockchain in cyberphysical systems, in 2020. His publications have attracted more than 9100 citations, with an h-index of 42. He serves on the editorial board of *Ad Hoc Networks*, *Nature Scientific Reports*, and on the organising and technical program committees of top international conferences, including Percom, ICBC, IPSN, WoWMoM, and ICDCS. He was TPC co-chair of ICBC, in 2021. He is a conjoint professor with the University of New South Wales, Australia.

Björn Schuller (Fellow, IEEE) received the diploma, in 1999, his doctoral degree for his study on automatic speech and emotion recognition, in 2006, and his habilitation and adjunct teaching professorship with the subject area of Signal Processing and Machine Intelligence, in 2012, all in electrical engineering and information technology from TUM in Munich/Germany. He is currently a professor of artificial intelligence with the Department of Computing, Imperial College London/U.K., where he heads GLAM – The Group on Language, Audio & Music, full professor and head of the chair of Embedded Intelligence for Health Care and Wellbeing with the University of Augsburg/Germany, and founding CEO/CSO of audEERING. He was previously full professor and head of the chair of Complex and Intelligent Systems with the University of Passau/Germany. He is Golden Core member of the IEEE Computer Society, senior member of the ACM, fellow of the BCS, fellow of the ISCA, president-emeritus and fellow of the Association for the Advancement of Affective Computing (AAAC), and was elected member of the IEEE Speech and Language Processing Technical Committee. He (co-)authored five books and more than 1 000 publications in peer-reviewed books, journals, and conference proceedings leading to more than overall 42 000 citations (h-index = 93). He was general chair of ACII 2019, co-program chair of Interspeech 2019 and ICMI 2019, repeated area chair of ICASSP, and former editor in chief of *IEEE Transactions on Affective Computing* next to a multitude of further associate and guest editor roles and functions in Technical and Organisational Committees.