Speaker-Invariant Adversarial Domain Adaptation for Emotion Recognition

Yufeng Yin
University of Southern California
Institute for Creative Technologies
Los Angeles, CA, USA
yin@ict.usc.edu

Baiyu Huang
University of Southern California
Institute for Creative Technologies
Los Angeles, CA, USA
baiyu@ict.usc.edu

Yizhen Wu
University of Southern California
Institute for Creative Technologies
Los Angeles, CA, USA
ywu@ict.usc.edu

Mohammad Soleymani
University of Southern California
Institute for Creative Technologies
Los Angeles, CA, USA
soleymani@ict.usc.edu

ABSTRACT

Automatic emotion recognition methods are sensitive to the variations across different datasets and their performance drops when evaluated across corpora. We can apply domain adaptation techniques *e.g.*, Domain-Adversarial Neural Network (DANN) to mitigate this problem. Though the DANN can detect and remove the bias between corpora, the bias between speakers still remains which results in reduced performance. In this paper, we propose Speaker-Invariant Domain-Adversarial Neural Network (SIDANN) to reduce both the domain bias and the speaker bias. Specifically, based on the DANN, we add a speaker discriminator to unlearn information representing speakers' individual characteristics with a gradient reversal layer (GRL). Our experiments with multimodal data (speech, vision, and text) and the cross-domain evaluation indicate that the proposed SIDANN outperforms (+5.6% and +2.8% on average for detecting arousal and valence) the DANN model, suggesting that the SIDANN has a better domain adaptation ability than the DANN. Besides, the modality contribution analysis shows that the acoustic features are the most informative for arousal detection while the lexical features perform the best for valence detection.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; **Artificial intelligence**; • **Human-centered computing**;

KEYWORDS

emotion recognition; domain adaptation; neural networks; multimodal learning

ICMI '20, October 25–29, 2020, Virtual event, Netherlands

ACM ISBN 978-1-4503-7581-8/20/10…$15.00
https://doi.org/10.1145/3382507.3418813

ACM Reference Format:
Yufeng Yin, Baiyu Huang, Yizhen Wu, and Mohammad Soleymani. 2020. Speaker-Invariant Adversarial Domain Adaptation for Emotion Recognition. In *Proceedings of the 2020 International Conference on Multimodal Interaction (ICMI '20), October 25–29, 2020, Virtual event, Netherlands.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3382507.3418813

1 INTRODUCTION

Emotions play a significant role not only in human creativity and intelligence but also in rational human thinking and decision-making. To enable natural and intelligent interaction with humans, computers need the ability to recognize and express emotions [24]. Over the past few years, deep learning approaches have shown promising performance for emotion recognition [34, 37, 42]. However, constructing a large-scale emotion benchmark is both time-consuming and expensive. As a result, it is unrealistic to construct a large fully-annotated database every time we perform an emotion recognition task on a new domain. Deep domain adaptation has emerged as a new learning technique to address the lack of massive amounts of labeled data [41]. Using the publicly available fully-annotated audiovisual emotion databases (*e.g.*, MSP-Improv [3], IEMOCAP [2]), we can apply deep domain adaptation techniques *e.g.*, DANN [9] to recognize the emotions on an unlabeled dataset.

In the adversarial-based domain adaptation *e.g.*, DANN [9], a domain discriminator is trained to classify whether a data point is drawn from the source or target domain. It is used to encourage the domain confusion through an adversarial objective to minimize the distance between the source and target domains [41]. The Domain-Adversarial Neural Network (DANN) [9] is trained to minimize the classification loss (for source samples) while maximizing domain confusion loss via the use of the GRL.

The DANN model succeeds in reducing the domain bias between the source and the target domains, but it fails to address the bias between the speakers. There are multiple speakers in the MSP-Improv and the IEMOCAP databases, each with their own individual appearance and voice characteristics. Though the DANN model can detect and remove the bias between domains, the bias between speakers still remains which results in reduced performance.

To address this problem, we propose Speaker-Invariant Domain-Adversarial Neural Network (SIDANN). Figure 1 shows the network

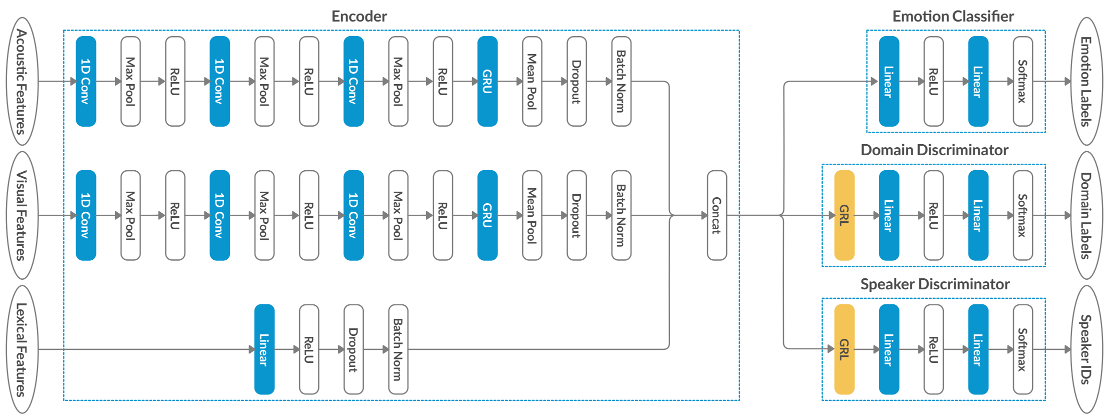


Figure 1: The network architecture for the cross-domain emotion recognition models. The inputs from different modalities are passed through the models, in this case, MFB/VGGish for speech, ResNet for vision, BERT for language. The baseline model only has the encoder and emotion classifier. The DANN model has the encoder, emotion classifier, and domain discriminator. The SIDANN has all the four parts (encoder, emotion classifier, domain discriminator, and speaker discriminator).

architectures for the cross-domain emotion recognition models. Specifically, based on the DANN model, we add a speaker discriminator to detect the speaker's identity. We add a GRL at the beginning of the discriminator so that the encoder can unlearn the speaker-specific information.

To confirm the effectiveness of our proposed model, we conduct the within-domain and cross-domain experiments with multimodal data (speech, vision, and text). We evaluate our method on two publicly available fully-annotated audiovisual emotion databases (MSP-Improv [3] and IEMOCAP [2]). Specifically, the MSP-Improv and IEMOCAP database have 8,348 and 10,039 utterances respectively produced by 22 speakers in total, each labeled with both arousal and valence values.

We extract two kinds of acoustic features: 1) the Mel Filter Bank (MFB) acoustic features. 2) the VGGish [10, 15] acoustic representations. We obtain the visual features from the penultimate layer of the ResNet-152 [14] and we use the pre-trained BERT [7] to transform the text from each utterance into a vector.

For the within-domain experiments, we train and test the baseline model with five-fold speaker-independent cross-validation. For the cross-domain experiments, we first train the baseline with the labeled source data and then train the domain adaptation models (DANN [9], and SIDANN) by fine-tuning the baseline with the labeled source data and unlabeled target data. We then test all three models on the whole target domain. The results of the within-domain experiments show that the multimodel with the MFB acoustic and the BERT lexical features has the best performance for arousal detection. Meanwhile, the multimodel with the MFB acoustic, the ResNet visual, and the BERT lexical features achieve the best performance for valence detection. For the cross-domain experiments, our results indicate that the proposed SIDANN model outperforms (+5.6% and +2.8% on average for detecting arousal and valence) the DANN model, confirming that the SIDANN has better domain adaptation ability than the DANN.

The major contributions of this work are as follows: (1) We study the unsupervised domain adaptation problem on emotion recognition with multimodal data including speech, vision, and language. We conduct detailed experiments to explore the domain adaptation performance of different modalities and their combinations. (2) We study the problem of how to reduce both the domain bias and speaker bias. Based on the DANN model, we propose Speaker-Invariant Domain-Adversarial Neural Network to separate the speaker bias from the domain bias. Specifically, we add a speaker discriminator to detect the speaker's identity. There is a GRL at the beginning of the discriminator so that the encoder can unlearn the speaker-specific information. The experimental results confirm that the SIDANN has a better domain adaptation ability than the DANN.

2 RELATED WORK

In this section, we introduce the background and the previous work of domain adaptation. Additionally, we show some research work applying domain adaptation techniques in emotion recognition.

2.1 Domain Adaptation

Supervised deep learning methods suffer from performance loss on unseen data due to the covariate shift. Domain adaptation techniques are proposed to reduce discrepancies between different domains. Unsupervised Domain Adaptation (UDA) can be used to train a model with labeled data from the source domain (training dataset) and unlabeled data from the target domain (unseen dataset). The goal is to learn a representation that is both discriminative for the main learning task (*e.g.*, emotion recognition) on the source domain and insensitive to the covariate shift between the domains.

Wang *et al.* [41] defines this kind of problem as the **homogeneous domain adaptation** and divides the homogeneous domain adaptation into three categories: discrepancy-based approach, adversarial-based approach, and reconstruction-based approach.

The discrepancy-based approach aims to diminish the shift between the two domains by fine-tuning the deep network model [41]. Tzeng *et al.* [36] proposes a new CNN architecture with an adaptation layer and an additional domain confusion loss, to learn a representation that is both semantically meaningful and domain invariant. Long *et al.* [20] proposes a Deep Adaptation Networks (DAN) architecture, which generalizes deep CNNs to the domain adaptation scenario. In this architecture, hidden representations of all task-specific layers are embedded in a reproducing kernel Hilbert space where the mean embeddings of different domain distributions can be explicitly matched. Rozantsev *et al.* [29] introduces a two-stream architecture, where one operates in the source domain and the other in the target domain. The weights in corresponding layers are related but not shared. Saito *et al.* [30] introduces a new approach that attempts to align distributions of source and target by utilizing the task-specific decision boundaries.

Regarding to the adversarial-based approach, a domain discriminator that classifies whether a data point is drawn from the source or target domain. It is used to encourage the domain confusion through an adversarial objective to minimize the distance between the source and target domains [41]. The Domain-Adversarial Neural Network (DANN) [9] integrates a gradient reversal layer (GRL) into the standard architecture to ensure that the feature distributions over the two domains are made similar. In contrast to the DANN, the Adversarial Discriminative Domain Adaptation (ADDA) [35] model considers the independent source and target mappings by untying the weights, and the parameters of the target model are initialized by the pre-trained source one. The Wasserstein Distance Guided Representation Learning (WDGRL) [32] uses a domain critic to minimize the Wasserstein Distance (with Gradient Penalty) between domains. The Multi-Adversarial Domain Adaptation (MADA) [23] captures multimode structures to enable fine-grained alignment of different data distributions based on multiple domain discriminators. The Selective Adversarial Network (SAN) [4] addresses partial transfer learning from big domains to small domains where the target label space is a subspace of the source label space.

The third category is the reconstruction-based approach which assumes that the data reconstruction of the source or target samples can help improve the performance of domain adaptation [41]. Bousmalis *et al.* [1] decouples domain adaptation from a specific task and trains a model that changes images from the source domain to appear as they were from the target domain while maintaining their original content. Hoffman *et al.* [16] proposes a novel discriminatively trained Cycle-Consistent Adversarial Domain Adaptation (CyCADA) model. The model adapts representations at both pixel- and feature-level and enforces cycle-consistency while leveraging a task loss, and does not require aligned pairs.

2.2 Domain Adaptation for Emotion Recognition

Because of the multi-faceted information included in the speech signal [8], domain adaptation has been widely applied to speech-based emotion recognition. Li *et al.* [19] proposes a machine learning framework to obtain speech emotion representations by limiting the effect of speaker variability in the speech signals. Gideon *et al.* [11] investigates how knowledge can be transferred between three paralinguistic tasks: speaker, emotion, and gender recognition.

Emotions result in behavioral changes including facial expressions [8]. A variety of domain adaptation techniques have been explored for vision-based emotion recognition. Zhao *et al.* [44] develops a novel adversarial model for emotion distribution learning, termed EmotionGAN, which optimizes the Generative Adversarial Network (GAN) loss, semantic consistency loss, and regression loss. The EmotionGAN model can adapt source domain images such that they appear as if they were drawn from the target domain while preserving the annotation information.

For cross-domain sentiment analysis, Glorot *et al.* [12] studies the problem of domain adaptation for sentiment classifiers. They demonstrated that a deep learning system based on Stacked Denoising Auto-Encoders with sparse rectifier units can perform an unsupervised feature extraction which is highly beneficial for the domain adaptation of sentiment classifiers.

Moreover, these modalities are often combined for multimodal learning. For example, Jaiswal *et al.* [17] studies how stress alters acoustic and lexical emotional detection. They use the GRL to decorrelate stress modulations from emotion representations. Zhao *et al.* [43] uses an adversarial training procedure to investigate how emotion knowledge of Western European cultures can be transferred to Chinese culture with all the three modalities (speech, vision, and language).

3 PROBLEM FORMULATION

Given a set of utterances S, for each utterance $\mathbf{x}_i \in S, \mathbf{x}_i = \{\mathbf{x}_i^a, \mathbf{x}_i^v, \mathbf{x}_i^l\}$, where $\mathbf{x}_i^a$, $\mathbf{x}_i^v$, and $\mathbf{x}_i^l$ represent the acoustic, visual, and lexical features respectively.

Problem. **Emotion Recognition.** Given an utterance set S, we aim to detect the arousal and the valence values a_i, v_i for each utterance $\mathbf{x}_i \in S$ using function $f_a(.)$ and $f_v(.)$:

$$a_i = f_a(\mathbf{x}_i) \tag{1}$$

$$v_i = f_v(\mathbf{x}_i) \tag{2}$$

4 DATASETS AND FEATURES

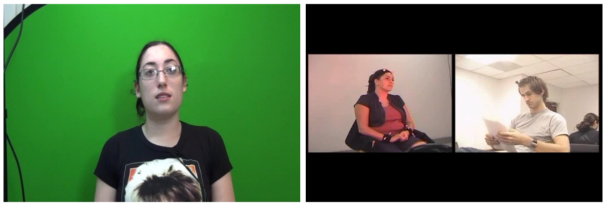
In this section, we introduce in detail the datasets we use to evaluate the methods.

4.1 Datasets

Two public datasets are used to study the UDA problem for emotion recognition: (1) MSP-Improv dataset [3]; and (2) Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [2]. Both are audiovisual databases and have the *arousal* and the *valence* labels. Videos from both the databases are shot in a laboratory thus they have similar environments.

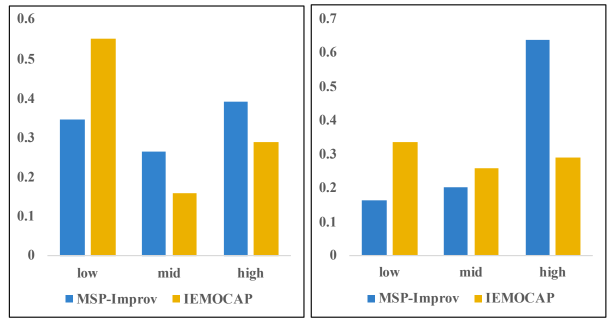
MSP-Improv. The MSP-Improv database is an acted audiovisual emotional database that explores emotional behaviors during acted and improvised dyadic interaction. Overall, the corpus consists of 8,438 turns (over 9 hours) of emotional sentences and 12 speakers (6 males and 6 females).

IEMOCAP. The IEMOCAP database is an acted, multimodal, and multispeaker database. It contains approximately 12 hours of audiovisual data, including video, speech, motion capture of face,



(a) A screenshot from MSP-Improv. (b) A screenshot from IEMOCAP.

Figure 2: Screenshots from MSP-Improv and IEMOCAP.



(a) Distributions of arousal values. (b) Distributions of valence values.

Figure 3: Label distributions for different domains.

text transcriptions. Overall, the dataset has 10,039 utterances and 10 speakers (5 males and 5 females).

Screenshots from the two databases are shown in Figure 2a and Figure 2b. Videos are recorded from different angles and the video resolutions are also different.

4.2 Labels

Each utterance in MSP-Improv and IEMOCAP has labels for both *arousal* and *valence* on a five-point Likert scale. According to the label processing method mentioned in [17], we bin the labels into one of the three classes, defined as, {"low":[1, 2.75], "mid":(2.75, 3.25], "high":(3.25, 5]}.

The overall distribution for *arousal* is: {"low": 45.69%, "mid": 20.72%, "high": 33.59%} and for *valence* is: {"low": 25.49%, "mid": 23.14%, "high": 51.37%}. Therefore, the label distributions are imbalanced. Moreover, label distributions vary between datasets (see Figure 3).

4.3 Features

Behavior from three modalities is analyzed. Speakers' spoken content is manually transcribed in the IEMOCAP and automatically recognized in the MSP-Improv. Videos are used to track facial expressions and speech prosody is analyzed from audio.

4.3.1 Speech. The Mel Filter Bank (MFB) consists of overlapping triangular filters with the cutoff frequencies determined by the center frequencies of the two adjacent filters [5]. The MFB acoustic features have shown great domain transferability in previous work of emotion recognition [17]. We use the same extraction method as [17]. Specifically, we extract the 40-dimensional MFB features using a 25-millisecond Hamming window with a step-size of 10-milliseconds. As a result, we have a $T \times 40$ vector for each utterance, where T represents the number of time steps.

Deep neural networks trained on large quantities of data are able to learn powerful representations [7, 10, 14, 15]. Therefore, we also utilize the VGGish [10, 15] to extract a deep generalized acoustic representation for different domains. VGGish is a deep convolutional neural network trained on audio spectrograms extracted from a large database of videos to recognize an ontology of 632 audio events, for example, vehicle noise, music genre, human locomotion [10, 15]. According to the acoustic feature extraction method mentioned in [33], we use the 128-dimensional embedding that can be generated by the VGGish after dimensionality reduction with Principal Component Analysis (PCA). We use the hop size of 33ms, which means a 128-dimensional vector is extracted for every 33ms of the audio signals. As a result, we have a $T \times 128$ vector for each utterance, where T represents the number of time steps.

4.3.2 Vision. We first sample the videos at a 30 fps rate and crop the speaker's face for each frame with OpenCV [1]. To extract a generalized visual representation for different domains, we extract the activations from the penultimate layer of ResNet-152 [14] trained on the ImageNet [6]. We feed the network with cropped faces from each frame. As a result, we have a $T \times 2048$ matrix for each utterance, where T denotes the number of frames.

4.3.3 Language. To represent the spoken words, we use the pre-trained BERT [7] for mapping the spoken utterances to a representation. Bidirectional Encoder Representations from Transformers (BERT) [7] is a method for learning a language model that can be trained on a large amount of data in an unsupervised manner. This pre-trained model is very effective in representing a sequence of terms as a fixed-length representation (vector). BERT representation achieves state-of-the-art results in multiple natural language understanding tasks [33]. In this paper, we use pre-trained BERT to transform the text for each utterance into a 768-dimensional vector. IEMOCAP includes manual transcriptions that we use for language analysis. We transcribe MSP-Improv using Google Cloud enhanced Automatic Speech Recognition (ASR) [2] to generate the text data. We discard 271 out of 8,438 utterances for which ASR fails to detect any speech. As a result, we use 8,167 utterances in total for the MSP-Improv database.

Finally, we z-normalize all the features from three modalities (acoustic, visual, and lexical) for each speaker, by subtracting their mean value and dividing them by their standard deviation.

5 METHODOLOGY

In this section, we introduce the notations and show the network architectures and the detailed training strategies for both the baseline and UDA models. Their network architectures are shown in Figure 1. Also, the pseudo-code for training the SIDANN model for one epoch is shown in the Algorithm 1.

[1]https://opencv.org/

[2]https://cloud.google.com/speech-to-text/docs/enhanced-models

5.1 Notations

Let the source dataset be $D_s = \{(\mathbf{x}_1, e_1, s_1, d_1), ..., (\mathbf{x}_M, e_M, s_M, d_M)\}$ and the target dataset be $D_t = \{(\mathbf{x}_{M+1}, s_{M+1}, d_{M+1}), ..., (\mathbf{x}_N, s_N, d_N)\}$ $(N > M > 0)$.

M and $N - M$ are the numbers of the source and target utterances respectively. $\mathbf{x}_i = \{\mathbf{x}_i^a, \mathbf{x}_i^v, \mathbf{x}_i^l\}$ is the extracted feature. e_i is the emotion label (arousal or valence value). We do not have the emotion labels for the target dataset. s_i denotes the speaker identity. d_i is the domain label, where $d_i = 0$ means x_i belongs to the source domain and $d_i = 1$ means it belongs to the target domain. Therefore, $d_i = 0$, for $i = 1, 2, ..., M$ and $d_i = 1$, for $i = M + 1, M + 2, ..., N$.

5.2 Baseline Model

We use the multimodal approach mentioned in [17]. It is worth noting that Jaiswal *et al.* [17] only utilizes the acoustic and lexical features to recognize emotions. Also, the lexical features they extracted are sequential but ours are not. Therefore, for the visual part, we use the same architecture as the acoustic one and for the lexical part, we simply use a linear layer as encoder.

The network architecture is shown in Figure 1. The baseline model only has two parts: encoder and emotion classifier. We assume each part as a mapping. The encoder G_e outputs a fixed-size representation $\mathbf{f}$ given $\mathbf{x}$ (acoustic, visual, and lexical features). The emotion classifier G_c maps $\mathbf{f}$ to a probability distribution $\mathbf{e}$ over the emotion label space of three classes (low or mid or high). We denote the vector of parameters from all layers in the encoder and the emotion classifier as θ_e and θ_c. As a result, we have:

$$\mathbf{f} = G_e(\mathbf{x}; \theta_e) \tag{3}$$

$$\mathbf{e} = G_c(\mathbf{f}; \theta_c) \tag{4}$$

The unimodal baseline only takes a single stream (acoustic or visual or lexical) input while the bimodal baseline takes a two-stream input and the trimodal baseline takes a three-steam input (acoustic, visual, and lexical).

The goal of the model is to minimize the cross-entropy loss which is defined as follows:

$$\begin{aligned} L_{Baseline} = L_{emotion} &= \sum_{(\mathbf{x}_i, e_i) \in D_s} L_e(G_c(\mathbf{f}_i; \theta_c), e_i) \\ &= \sum_{(\mathbf{x}_i, e_i) \in D_s} L_e(G_c(G_e(\mathbf{x}_i; \theta_e); \theta_c), e_i) \end{aligned} \tag{5}$$

Where L_e is the cross-entropy loss.

5.3 Domain-Adversarial Neural Network

The Domain-Adversarial Neural Network (DANN) [9] minimizes the classification loss (for source samples) while maximizing domain confusion loss. The DANN integrates a gradient reversal layer (GRL) into the standard architecture to ensure that the feature distributions over the two domains are similar.

Based on the baseline architecture, we add a domain discriminator to discriminate whether the output of the encoder is from the source or the target domain. Specifically, there is a gradient reversal layer (GRL) at the beginning of the domain discriminator.

The DANN has three parts: encoder, emotion classifier, and domain discriminator. The domain discriminator G_d maps $\mathbf{f}$ to a probability distribution $\mathbf{d}$ over the domain label space of two classes (source or target). We denote the vector of parameters from all layers in the domain discriminator as θ_d. Therefore, we have:

$$\mathbf{d} = G_d(\mathbf{f}; \theta_d) \tag{6}$$

The objective function of the model has two parts: the task-specific loss and domain loss. The task-specific loss is the same as the baseline objective function which is shown in Equation 5. The domain loss is defined as follow:

$$\begin{aligned} L_{domain} &= \sum_{(\mathbf{x}_i, d_i) \in D_s \cup D_t} L_d(G_d(\mathbf{f}_i; \theta_d), d_i) \\ &= \sum_{(\mathbf{x}_i, d_i) \in D_s \cup D_t} L_d(G_d(G_e(\mathbf{x}_i; \theta_e); \theta_d), d_i) \end{aligned} \tag{7}$$

Where L_d is the cross-entropy loss.

The objective of the DANN is to maximize the performance of the emotion classifier while minimizing the performance of the domain discriminator. Overall, the goal of the DANN model is defined as follows:

$$L_{DANN} = L_{emotion} - \lambda * L_{domain} \tag{8}$$

Where λ is the hyper-parameter that controls the trade-off between the two objectives that shape the features during learning [9].

Algorithm 1 Train the SIDANN for one epoch. For Adam optimizer, we use the default values of $\alpha = 0.0001$, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. The batch size m is 256.

Require: The batch size m, Adam hyperparameters α, β_1, β_2.
Require: Parameters for encoder θ_e, emotion classifier θ_c, domain discriminator θ_d, and speaker discriminator θ_s and their corresponding mappings: G_c, G_e, G_d, and G_s.
Require: Weights for the domain loss λ_1 and the speaker loss λ_2.
$m' \leftarrow m/2$
$n_1 \leftarrow$ (Number of source samples)$/m'$
$n_2 \leftarrow$ (Number of target samples)$/m'$
$n \leftarrow \min(n_1, n_2)$
for $batch = 1, ..., n$ **do**
Sample $\{\mathbf{x}_i, e_i, s_i, d_i\}_{i=1}^{m'}$ a half batch from source data
Sample $\{\mathbf{x}_i, s_i, d_i\}_{i=m'+1}^{m}$ a half batch from target data
$\mathbf{X}_s \leftarrow \{\mathbf{x}_i\}_{i=1}^{m'}$
$\mathbf{X} \leftarrow \{\mathbf{x}_i\}_{i=1}^{m}$
$\mathbf{f}_s \leftarrow G_e(\mathbf{X}_s)$
$\mathbf{f} \leftarrow G_e(\mathbf{X})$
$L^e \leftarrow G_e(\mathbf{f}_s)$
$L^d \leftarrow G_d(\mathbf{f})$
$L^s \leftarrow G_s(\mathbf{f})$
$loss_E \leftarrow \frac{1}{m'} \sum_{i=1}^{m'} e_i \times \log(L_i^e)$
$loss_D \leftarrow \frac{1}{m} \sum_{i=1}^{m} d_i \times \log(L_i^d)$
$loss_S \leftarrow \frac{1}{m} \sum_{i=1}^{m} s_i \times \log(L_i^s)$
$\theta_e \leftarrow \text{Adam}(\Delta_{\theta_e}[loss_E - \lambda_1 loss_D - \lambda_2 loss_S], \theta_e, \alpha, \beta_1, \beta_2)$
$\theta_c \leftarrow \text{Adam}(\Delta_{\theta_c}[loss_E], \theta_c, \alpha, \beta_1, \beta_2)$
$\theta_d \leftarrow \text{Adam}(\Delta_{\theta_d}[loss_D], \theta_d, \alpha, \beta_1, \beta_2)$
$\theta_s \leftarrow \text{Adam}(\Delta_{\theta_s}[loss_S], \theta_s, \alpha, \beta_1, \beta_2)$
end for

Table 1: Within-domain performance of the baseline model. A1, A2, V, and L represent VGGish acoustic, MFB acoustic, ResNet visual, and BERT lexical features. M and I stand for MSP-Improv and IEMOCAP database. ACC and UAR stand for Accuracy and Unweighted Average Recall. They are 0.33 when the detected labels are uniformly distributed.

(a) Results for detecting arousal.

	ACC		UAR		
Modality	M	I	M	I	Avg
A1	0.619	0.577	0.509	0.525	0.558
A2	0.672	**0.593**	0.492	**0.602**	0.590
V	0.568	0.474	0.415	0.492	0.487
L	0.513	0.510	0.409	0.473	0.476
A1+V	0.601	0.569	0.455	0.532	0.539
A2+V	0.646	0.556	0.503	0.489	0.549
A1+L	0.587	0.578	0.467	0.527	0.540
A2+L	**0.684**	0.587	**0.587**	0.550	**0.602**
V+L	0.547	0.505	0.428	0.466	0.487
A1+V+L	0.623	0.573	0.513	0.517	0.557
A2+V+L	0.644	0.570	0.500	0.530	0.561

(b) Results for detecting valence.

	ACC		UAR		
Modality	M	I	M	I	Avg
A1	0.428	0.466	0.417	0.431	0.436
A2	0.422	0.455	0.489	0.489	0.464
V	0.503	0.499	0.472	0.462	0.484
L	0.513	0.618	0.499	0.576	0.552
A1+V	0.503	0.518	0.493	0.462	0.494
A2+V	0.489	0.510	0.471	0.478	0.487
A1+L	0.538	0.611	0.515	0.571	0.559
A2+L	0.538	0.629	0.527	**0.583**	0.569
V+L	0.539	0.614	0.520	0.541	0.554
A1+V+L	**0.554**	**0.643**	0.534	0.555	0.572
A2+V+L	0.537	0.638	**0.553**	0.571	**0.575**

5.4 Speaker-Invariant Domain-Adversarial Neural Network

Although the DANN model can remove the domain bias between the source and the target domain, it ignores the bias between speakers. There are 12 speakers in the MSP-Improv database and 10 in the IEMOCAP database. These 22 speakers have individual styles for expressing emotions. Therefore, during the DANN training, the model mixes these two sources of bias together resulting in poor performance.

To address this problem, we propose Speaker-Invariant Domain-Adversarial Neural Network (SIDANN). Specifically, we add a speaker discriminator to detect the speaker's identity. Similar to the DANN model, we add a GRL at the beginning of the discriminator so that the encoder can unlearn the speaker-specific information. With the speaker discriminator, the model can separate the speaker bias from the domain bias.

Overall, the SIDANN has four parts: encoder, emotion classifier, domain discriminator, and speaker discriminator. The speaker discriminator G_s maps $\mathbf{f}$ to a probability distribution $\mathbf{s}$ over the speaker label space of 22 classes. We denote the vector of parameters from all layers in the speaker discriminator as θ_s. Therefore, we have:

$$\mathbf{s} = G_s(\mathbf{f}; \theta_s) \tag{9}$$

Besides the task-specific loss (Equation 5) and domain loss (Equation 7), the objective function of the SIDANN has the speaker loss, which is defined as follows:

$$\begin{aligned} L_{speaker} &= \sum_{(\mathbf{x}_i, s_i) \in D_s \cup D_t} L_s(G_s(\mathbf{f}_i; \theta_s), s_i) \\ &= \sum_{(\mathbf{x}_i, s_i) \in D_s \cup D_t} L_s(G_s(G_e(\mathbf{x}_i; \theta_e); \theta_s), s_i) \end{aligned} \tag{10}$$

Where L_s is the cross-entropy loss.

The objective of the SIDANN is to maximize the performance of the emotion classifier while minimizing the performance of the domain discriminator and the speaker discriminator.

Integrating all the things (Equation 5, 7, and 10), the goal of the DANN model is defined as follows:

$$L_{SIDANN} = L_{emotion} - \lambda_1 * L_{domain} - \lambda_2 * L_{speaker} \tag{11}$$

Where λ_1 and λ_2 are the hyperparameters that control the trade-off between the three objectives that shape the features during learning.

The pseudo-code for training the SIDANN model for one epoch is shown in the Algorithm 1.

6 EXPERIMENTS

In this section, we will describe the experimental design and the training details. We will also report and discuss the experimental results.

6.1 Experimental Design

6.1.1 Within-domain Evaluation. To evaluate the baseline model, we train and test it with five-fold speaker-independent cross-validation. Specifically, we evaluate the performance of the unimodal, bimodal, and trimodal model.

6.1.2 Cross-domain Evaluation. We design to set one database as the source domain and the other as the target domain. Thus, we have two directions of domain adaptation (M $\rightarrow$ I and I $\rightarrow$ M, where M is MSP-Improv and I is IEMOCAP).

For the baseline model, we use 80% of the source data for training and 20% for validation where training and validation data are speaker-independent. For the DANN and the SIDANN, we train them by fine-tuning the baseline model with the labeled source data and unlabeled target data. We then test all the three models on the whole target domain.

6.2 Evaluation Metrics

We utilize Accuracy (ACC) and Unweighted Average Recall (UAR) to evaluate the performance. Specifically, ACC and UAR are 0.33 when the detected labels are uniformly distributed.

Table 2: Cross-domain performance of the unsupervised domain adaptation.

(a) Results for detecting arousal (Inputs are the MFB acoustic features and the BERT lexical features).

	ACC		UAR		
Model	M → I	I → M	M → I	I → M	Avg
Baseline	0.241(.03)	0.291(.05)	0.186(.02)	0.245(.03)	0.241
DANN	0.321(.01)	0.266(.06)	0.271(.01)	0.279(.02)	0.309
SIDANN	**0.392(.03)**	**0.390(.08)**	**0.371(.02)**	**0.308(.01)**	**0.365**

(b) Results for detecting arousal (Inputs are the MFB acoustic features and the ResNet visual features).

	ACC		UAR		
Model	M → I	I → M	M → I	I → M	Avg
Baseline	0.263(.02)	0.284(.06)	0.188(.01)	0.277(.03)	0.253
DANN	0.388(.03)	0.407(.09)	0.344(.02)	0.336(.03)	0.369
SIDANN	**0.415(.01)**	**0.506(.07)**	**0.422(.03)**	**0.379(.03)**	**0.430**

(c) Results for detecting valence (Inputs are the MFB acoustic features, the ResNet visual features, and the BERT lexical features).

	ACC		UAR		
Model	M → I	I → M	M → I	I → M	Avg
Baseline	0.381(.02)	0.407(.01)	0.442(.02)	0.406(.01)	0.409
DANN	0.460(.02)	0.456(.02)	0.409(.01)	0.456(.03)	0.445
SIDANN	**0.480(.01)**	**0.500(.03)**	**0.431(.02)**	**0.482(.03)**	**0.473**

6.3 Training Details

For the baseline model, it is trained for a maximum of 50 epochs and we stop the training if the validation loss does not improve after five consecutive epochs. Given the imbalanced nature of our data, we utilize an imbalanced dataset sampler [3] to re-balance the training class distributions. The model is trained with the Adam [18] optimizer (initial learning rate = 10^{-4}) with a dynamic learning rate decay [4] based on the validation loss. We use the default parameters for the Adam optimizer. The batch size is 256. All models are implemented in PyTorch [22].

We use validation samples (20% source data) for hyper-parameter selection and early stopping. The hyperparameters that we use for the baseline include: the width of the convolution layers {64, 128}, the kernel size of the convolution layers {2, 3}, the kernel size of the max pool layers {2}, the number of the GRU layers {2, 3}, the width of the linear layer in encoder {32}, the width of the linear layer in emotion classifier {32, 64}, and the dropout rate {0.3}.

For the UDA models, they are simply trained for 25 epochs, since we do not have the labels for the target domain. They are trained with the Adam optimizer with a fixed learning rate, which is also a hyper-parameter. The optimizer is set with the default parameters. The batch size is also 256.

The network structures of the domain discriminator and the speaker discriminator are exactly the same as that of the emotion classifier. The hyperparameters we use for the DANN include: the learning rate {1e-5, 3e-5, 1e-4, 3e-4, 1e-3}, and λ {0.1, 0.3, 1, 3, 10} while for the SIDANN include: the learning rate {1e-5, 3e-5, 1e-4, 3e-4, 1e-3}, λ_1 {0.1, 0.3, 1, 3, 10}, and λ_2 {0.1, 0.3, 1, 3, 10}, where the meanings of λ, λ_1, λ_2 have been explained in Section 5.3 and Section 5.4.

[3]https://github.com/ufoym/imbalanced-dataset-sampler

[4]https://pytorch.org/docs/stable/optim.html#torch.optim.lr_scheduler.ReduceLROnPlateau

Table 3: Cross-domain performance with the MFB acoustic features.

(a) Results for detecting arousal.

	ACC		UAR		
Model	M → I	I → M	M → I	I → M	Avg
Baseline	0.258	0.368	0.201	0.224	0.263
DANN	0.367	0.414	0.445	0.306	0.383
SIDANN	**0.407**	**0.496**	**0.452**	**0.428**	**0.446**
Jaiswal *et al.* [17]	-	-	-	0.402	-

(b) Results for detecting valence.

	ACC		UAR		
Model	M → I	I → M	M → I	I → M	Avg
Baseline	0.464	0.367	0.402	0.364	0.399
DANN	0.470	0.376	0.442	0.406	0.424
SIDANN	**0.550**	**0.407**	**0.482**	**0.452**	**0.473**
Jaiswal *et al.* [17]	-	-	-	0.439	-

6.4 Experimental Results

6.4.1 Within-domain Evaluation. Table 1 displays the within-domain performances of the baseline model. We have totally evaluated 11 models of different feature combinations.

For arousal detection (shown in Table 1a), the MFB combined with the BERT features has the best ACC scores while the MFB features achieve the highest UAR scores on both the MSP-Improv and IEMOCAP databases. Further, the MFB combined with the BERT features works better than the MFB features on average. For unimodal methods, both acoustic features perform better than the other modalities (vision and language) and the lexical features perform the worst on average. For valence detection (shown in Table 1b), the VGGish combined with the ResNet and BERT features achieve the best ACC scores on both the MSP-Improv and IEMOCAP databases. However, the MFB combined with the ResNet and BERT features has the best performance on average. For unimodal methods, lexical features perform the best while acoustic features are the worst. This is the exact opposite of arousal detection. The acoustic features are informative for arousal detection and the lexical features are powerful for valence detection. Past work [13, 21] showed that speech works better for arousal detection and language is better able to capture valence. Facial expression is also better at detecting valence than arousal, see AVEC challenges results [25–28, 31, 38–40].

6.4.2 Cross-domain Evaluation. We show the results of the cross-domain performance in Table 2. We input the MFB and the BERT features for detecting arousal and the MFB, the ResNet, and the BERT features for detecting valence since these two combinations have the highest performance on average for each task. We report

Table 4: Modality contribution analysis for unsupervised domain adaptation.

(a) Results for detecting arousal.

	VGGish acoustic features				ResNet visual features				BERT lexical features				
	ACC		UAR		ACC		UAR		ACC		UAR		
Model	M → I	I → M	M → I	I → M	M → I	I → M	M → I	I → M	M → I	I → M	M → I	I → M	Avg
Baseline	0.311	0.360	0.244	0.300	0.405	0.280	0.340	0.297	0.276	0.317	0.259	0.290	0.307
DANN	**0.503**	**0.559**	0.467	0.353	**0.453**	0.430	0.413	0.367	0.313	0.323	0.304	0.310	0.400
SIDANN	0.491	0.556	**0.485**	**0.376**	0.415	**0.437**	**0.485**	**0.378**	**0.315**	**0.327**	**0.309**	**0.313**	**0.407**

(b) Results for detecting valence.

	VGGish acoustic features				ResNet visual features				BERT lexical features				
	ACC		UAR		ACC		UAR		ACC		UAR		
Model	M → I	I → M	M → I	I → M	M → I	I → M	M → I	I → M	M → I	I → M	M → I	I → M	Avg
Baseline	0.424	0.360	0.376	0.386	0.323	0.350	0.343	0.316	0.494	0.453	0.475	0.449	0.396
DANN	0.450	**0.401**	0.395	0.402	0.499	0.400	0.459	0.381	0.506	0.458	**0.484**	0.458	0.441
SIDANN	**0.481**	0.400	**0.414**	**0.405**	**0.501**	**0.408**	**0.510**	**0.400**	**0.515**	**0.467**	0.477	**0.465**	**0.454**

the results in Table 2a and Table 2c. The numbers in the brackets are the standard deviations. The numbers indicate that our proposed model performs significantly better than the DANN and the baseline with t-test (at $p < 0.1$). Specifically, the SIDANN outperforms the DANN by 5.6% and 2.8% on average for detecting arousal and valence, confirming that the SIDANN has a better domain adaptation ability than the DANN.

Though the SIDANN is the best performing model, it performs poorly detecting arousal. Based on the modality contribution analysis in Section 6.5, we speculate that the lexical features are not helpful for detecting arousal. Therefore, we replace the BERT lexical features with the ResNet visual features and display the results in Table 2b. The results show that the MFB combined the ResNet features work better than the MFB combined with the BERT features for all the evaluation metrics. Specifically, the former one outperforms the later one by 6.1% on average. The result is significant at $p < 0.1$ with t-test.

6.5 Modality Contribution Analysis

To figure out the contribution of each modality, we re-conduct the cross-domain experiment with a single modality (acoustic or visual or lexical). Specifically, we first train unimodal models on the source domain and then fine-tune them. The results of the modality contribution analysis are reported in Table 3 and Table 4.

Table 3 shows the cross-domain performance with the MFB acoustic features. The proposed SIDANN model performs better than the DANN and baseline model for both arousal and valence. Specifically, the SIDANN outperforms the DANN by 6.3% and 4.9% on average when detecting arousal and valence values respectively. Also, the proposed model achieves higher UAR than the numbers reported in [17]. The results of the other three kinds of features (VGGish, ResNet, and BERT) are reported in Table 4. The SIDANN has a slight advantage over the DANN (+0.7% and +1.3% for arousal and valence). Additionally, we find that the BERT lexical features perform worst for arousal detection while they perform best for valence detection. This is consistent with the previous results we obtain in the within-domain experiments.

7 CONCLUSIONS

In this work, we study the Unsupervised Domain Adaptation (UDA) problem on emotion recognition with multimodal data including speech, vision, and language. We propose Speaker-Invariant Domain-Adversarial Neural Network (SIDANN) to separate the speaker bias from the domain bias. Specifically, we add a speaker discriminator to detect the speaker's identity. There is a gradient reversal layer at the beginning of the discriminator so that the encoder can unlearn the speaker-specific information. The cross-domain experimental results indicate that the proposed SIDANN model outperforms (+5.6% and +2.8% on average for detecting arousal and valence) the DANN model, confirming that the SIDANN has a better domain adaptation ability than the DANN.

Though the multimodal methods perform better than the unimodal methods for the within-domain experiments, the results of later ones are better for the cross-domain experiments. Therefore, for our future work, we need to explore additional multimodal fusion techniques to solve the problem. We also plan to evaluate our proposed model on other tasks to evaluate its general ability to reduce between-subject variance.

ACKNOWLEDGMENTS

Research was sponsored by the Army Research Office and was accomplished under Cooperative Agreement Number W911NF-20-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

REFERENCES

[1] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. 2017. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3722–3731.

[2] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation* 42, 4 (2008), 335.

[3] Carlos Busso, Srinivas Parthasarathy, Alec Burmania, Mohammed AbdelWahab, Najmeh Sadoughi, and Emily Mower Provost. 2016. MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception. *IEEE Transactions on Affective Computing* 8, 1 (2016), 67–80.

[4] Zhangjie Cao, Mingsheng Long, Jianmin Wang, and Michael I Jordan. 2018. Partial transfer learning with selective adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2724–2732.

[5] Yashpalsing Chavhan, ML Dhore, and Pallavi Yesaware. 2010. Speech emotion recognition using support vector machine. *International Journal of Computer Applications* 1, 20 (2010), 6–9.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 4171–4186.

[8] Kexin Feng and Theodora Chaspari. 2020. A Review of Generalizable Transfer Learning in Automatic Emotion Recognition. *Frontiers in Computer Science* 2 (2020), 9.

[9] Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised Domain Adaptation by Backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Francis Bach and David Blei (Eds.), Vol. 37. PMLR, Lille, France, 1180–1189. http://proceedings.mlr.press/v37/ganin15.html

[10] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 776–780.

[11] John Gideon, Soheil Khorram, Zakaria Aldeneh, Dimitrios Dimitriadis, and Emily Mower Provost. 2017. Progressive Neural Networks for Transfer Learning in Emotion Recognition. *Proc. Interspeech 2017* (2017), 1098–1102.

[12] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. (2011).

[13] Hatice Gunes and Björn Schuller. 2013. Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image and Vision Computing* 31, 2 (2013), 120–136.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[15] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. 2017. CNN architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 131–135.

[16] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. 2018. CyCADA: Cycle-Consistent Adversarial Domain Adaptation. In *Proceedings of the 35th International Conference on Machine Learning*.

[17] Mimansa Jaiswal, Zakaria Aldeneh, and Emily Mower Provost. 2019. Controlling for Confounders in Multimodal Emotion Classification via Adversarial Learning. In *2019 International Conference on Multimodal Interaction*. 174–184.

[18] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[19] Haoqi Li, Ming Tu, Jing Huang, Shrikanth Narayanan, and Panayiotis Georgiou. 2020. Speaker-Invariant Affective Representation Learning via Adversarial Training. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7144–7148.

[20] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. 2015. Learning transferable features with deep adaptation networks. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning-Volume 37*. JMLR. org, 97–105.

[21] Mihalis A Nicolaou, Hatice Gunes, and Maja Pantic. 2011. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Transactions on Affective Computing* 2, 2 (2011), 92–105.

[22] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. In *NeurIPS Autodiff Workshop*.

[23] Zhongyi Pei, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. 2018. Multi-adversarial domain adaptation. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[24] Rosalind W Picard. 2000. *Affective computing*. MIT press.

[25] Fabien Ringeval, Björn Schuller, Michel Valstar, Roddy Cowie, Heysem Kaya, Maximilian Schmitt, Shahin Amiriparian, Nicholas Cummins, Denis Lalanne, Adrien Michaud, et al. 2018. AVEC 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition. In *Proceedings of the 2018 on audio/visual emotion challenge and workshop*. 3–13.

[26] Fabien Ringeval, Björn Schuller, Michel Valstar, Roddy Cowie, and Maja Pantic. 2015. AVEC 2015: The 5th international audio/visual emotion challenge and workshop. In *Proceedings of the 23rd ACM international conference on Multimedia*. 1335–1336.

[27] Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, Shahin Amiriparian, Eva-Maria Messner, et al. 2019. AVEC 2019 workshop and challenge: state-of-mind, detecting depression with AI, and cross-cultural affect recognition. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*. 3–12.

[28] Fabien Ringeval, Björn Schuller, Michel Valstar, Jonathan Gratch, Roddy Cowie, Stefan Scherer, Sharon Mozgai, Nicholas Cummins, Maximilian Schmitt, and Maja Pantic. 2017. Avec 2017: Real-life depression, and affect recognition workshop and challenge. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. 3–9.

[29] Artem Rozantsev, Mathieu Salzmann, and Pascal Fua. 2018. Beyond sharing weights for deep domain adaptation. *IEEE transactions on pattern analysis and machine intelligence* 41, 4 (2018), 801–814.

[30] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. 2018. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3723–3732.

[31] Björn Schuller, Michel Valster, Florian Eyben, Roddy Cowie, and Maja Pantic. 2012. AVEC 2012: the continuous audio/visual emotion challenge. In *Proceedings of the 14th ACM international conference on Multimodal interaction*. 449–456.

[32] Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. 2018. Wasserstein distance guided representation learning for domain adaptation. In *Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*.

[33] Mohammad Soleymani, Kalin Stefanov, Sin-Hwa Kang, Jan Ondras, and Jonathan Gratch. 2019. Multimodal Analysis and Estimation of Intimate Self-Disclosure. In *2019 International Conference on Multimodal Interaction*. 59–68.

[34] George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A Nicolaou, Björn Schuller, and Stefanos Zafeiriou. 2016. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 5200–5204.

[35] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. 2017. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7167–7176.

[36] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. 2014. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474* (2014).

[37] Panagiotis Tzirakis, George Trigeorgis, Mihalis A Nicolaou, Björn W Schuller, and Stefanos Zafeiriou. 2017. End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of Selected Topics in Signal Processing* 11, 8 (2017), 1301–1309.

[38] Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. 2016. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th international workshop on audio/visual emotion challenge*. 3–10.

[39] Michel Valstar, Björn Schuller, Kirsty Smith, Timur Almaev, Florian Eyben, Jarek Krajewski, Roddy Cowie, and Maja Pantic. 2014. Avec 2014: 3d dimensional affect and depression recognition challenge. In *Proceedings of the 4th international workshop on audio/visual emotion challenge*. 3–10.

[40] Michel Valstar, Björn Schuller, Kirsty Smith, Florian Eyben, Bihan Jiang, Sanjay Bilakhia, Sebastian Schnieder, Roddy Cowie, and Maja Pantic. 2013. AVEC 2013: the continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*. 3–10.

[41] Mei Wang and Weihong Deng. 2018. Deep visual domain adaptation: A survey. *Neurocomputing* 312 (2018), 135–153.

[42] Shiqing Zhang, Shiliang Zhang, Tiejun Huang, Wen Gao, and Qi Tian. 2017. Learning affective features with a hybrid deep model for audio–visual emotion recognition. *IEEE Transactions on Circuits and Systems for Video Technology* 28, 10 (2017), 3030–3043.

[43] Jinming Zhao, Ruichen Li, Jingjun Liang, Shizhe Chen, and Qin Jin. 2019. Adversarial Domain Adaption for Multi-Cultural Dimensional Emotion Recognition in Dyadic Interactions. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*. 37–45.

[44] Sicheng Zhao, Xin Zhao, Guiguang Ding, and Kurt Keutzer. 2018. EmotionGAN: Unsupervised domain adaptation for learning discrete probability distributions of image emotions. In *Proceedings of the 26th ACM international conference on Multimedia*. 1319–1327.