This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2023.3290795

Adversarial Domain Generalized Transformer for Cross-Corpus Speech Emotion Recognition

Yuan Gao, Longbiao Wang, Jiaxing Liu, Jianwu Dang, and Shogo Okada

Abstract—Speech emotion recognition (SER) promotes the development of intelligent devices, which enable natural and friendly human-computer interactions. However, the recognition performance of existing approaches is significantly reduced on unseen datasets, and the lack of sufficient training data limits the generalizability of deep learning models. In this work, we analyze the impact of the domain generalization method on cross-corpus SER and propose an adversarial domain generalized transformer (ADoGT), which is aimed at learning a shared feature distribution for the source and target domains. Specifically, we investigate the effect of domain adversarial learning by eliminating nonaffective information. We also combine the center loss with the softmax function as joint supervision to learn discriminative features. Moreover, we introduce unsupervised transfer learning to extract additional features, and incorporate a gated fusion model to learn the complementary information of the features learned by the supervised feature extractor and pretrained model. The proposed transformer based domain generalization method is evaluated using four emotional datasets. We also provide an ablation study of different domain adversarial model structures and feature fusion models. The results of comparative experiments demonstrate the effectiveness of the proposed ADoGT.

Index Terms—Speech emotion recognition, cross-corpus, adversarial learning, domain generalization.

1 Introduction

HUMAN-computer interactions have become pervasive in our daily lives, and understanding human emotion is crucial for the development of intelligent devices [1]. Therefore, research on sentiment analysis and emotion recognition has attracted increasing attention in both industry and academia [2]. Speech emotion recognition (SER) is aimed at identifying emotional attributes in human speech, and a robust SER system can promote the development of empathetic chatbots and enrich the manual service of call centers [3]. This research also has other applications, such as monitoring the attention status of students in online courses, tracking the emotional state of patients with depression and providing advice about their diagnoses [4]. Previous studies designed empirical low-level descriptors (LLDs) for emotion classification [5]. In recent years, some researchers have found that deep learning based models such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) show promising results in SER tasks without expert knowledge [6], [7], [8].

Despite the recent progress in SER research, two bottlenecks limit the recognition accuracy of existing cross-corpus approaches. The first bottleneck is the lack of sufficient labeled training data [9]. Compared with other speech signal processing tasks such as automatic speech recognition, collecting and annotating speech data with emotional labels in natural environments is time-consuming. The number of utterances in most emotional datasets is not sufficient to train robust deep learning models [10]. The second bottleneck is how to extract discriminative features from speech signals. As human emotion is sometimes ambiguous, extracting optimal features from acoustic signals requires considerable attention [11]. Moreover, in cross-corpus evaluations, the emotional information in speech is difficult to learn due to variations in the domain information [12]. Because of this domain divergence, common deep learning models show poor performance on unseen datasets [13]. Most of the existing approaches are trained and tested with the same dataset, and the performance is significantly reduced on unseen datasets [14].

To address the lack of annotated emotional data, we use two types of feature extractors: 1) CNNs have shown promising performance in extracting emotional discriminative features for SER. Thus, we use a deep CNN architecture as the feature extractor in our baseline system to learn the spatial information of input utterances. 2) We pretrain an unsupervised convolutional autoencoder to transfer prior knowledge and extract bottleneck features as additional inputs for emotion classification. In this study, to improve the generalizability of the SER system, we propose the adversarial domain generalized Transformer (ADoGT), which effectively reduces the domain divergence between the training and test data and obtains more effective feature representations for each input utterance. Previous studies have identified that emotional information can be lost after feature compression [15]. Our proposed Transformer based feature encoder can retain sufficient emotional information from different feature distributions and achieve dimension reduction through multihead attention [16]. Furthermore, we incorporate a gated fusion model with our feature extractor to learn the complementary information in the two branches of the feature extractor. To address domain mis-

- *Y. Gao, L. Wang, J. Liu, and J Dang are with the Tianjin Key Laboratory of Cognitive Computing and Application, College of Intelligence and Computing, Tianjin University, Tianjin, China. E-mail: {yuan_gao, longbiao_wang, jiaxing_liu, dangjianwu}@tju.edu.cn (Corresponding author: L. Wang)*
- *S. Okada is with the School of Information Science, Japan Advanced Institute of Science and Technology (JAIST), Ishikawa, Japan. E-mail: {okada-s}@jaist.ac.jp*

match issues, researchers often use the adversarial domain adaptation method to transfer the domain representation from the source domain to the target domain [17]. In this study, we incorporate a domain adversarial neural network (DANN) to eliminate the speaker, corpus, and other domain information of the latent representation. Domain adaptation is achieved by reversing the gradient between the feature extractor and the domain classifier [18], which enables our model to maximize the training loss of nonaffective information. Moreover, in previous works, emotion classifiers commonly used the softmax loss function to identify decision boundaries and separate different emotions [19], [20]. We incorporate the center loss [21] with an emotion classifier to learn more cohesive features for SER. Therefore, the proposed model can learn a shared feature distribution for the source and target domains and thus achieve domain generalization in cross-corpus SER tasks.

The main contributions of this paper can be summarized as follows: (1) We address the domain divergence in cross-corpus SER by the proposed domain generalization method, which combines domain adversarial learning and center loss to generalize the feature distributions of different domains. (2) Our model incorporates the gated fusion model with the Transformer encoder to effectively combine the feature representation of the supervised and unsupervised feature encoder. (3) We analyze the domain generalization performance in addressing language mismatch issues and different elicitation types to meet real-world scenarios. (4) We explore the impact of different DANN subtasks in multi-domain SER and compare the domain classifier in the DANN with multi-task learning classifiers to analyze the effect of domain adversarial learning. The remainder of this paper is organized as follows: We provide a literature review of cross-corpus SER in Section 2. Then, we describe the details of our proposed algorithm in Section 3. The emotional datasets and experimental settings are presented in Section 4. In Section 5, we provide comparative experiments to evaluate the effectiveness of our model. We conclude this paper and outline our future work in Section 6.

2 Related Works

2.1 Cross-Corpus Speech Emotion Recognition

In real-world scenarios, several paralingual factors impact the acoustic features of speech signals, making it difficult for common machine learning models to learn emotional information in speech [22]. The mismatches between different datasets affect the performance of existing SER systems. Domain mismatch has various causes, including the language, recording conditions, and elicitation methods [12]. Another problem for SER is data sparsity. Since recording and annotating emotional speech is time consuming, the training data are often not sufficient to build robust SER systems. Compared with other speech signal processing tasks such as speech recognition, the limited data in SER tasks worsens the domain divergence problem [23]. Moreover, the ground truth cannot be objectively defined since the emotional labels are derived from perceptual evaluations and usually vary among annotators. To improve the generalizability of SER, researchers have focused on cross-corpus and multi-corpus evaluations [24], [25], [26].

In [12], Schuller et al. selected six existing datasets to explore the impact of the feature selection strategy and address different emotion annotations in cross-corpus SER tasks. To address mismatched acoustic conditions between the training and test data, the authors investigated several normalization methods, including speaker, corpus, and speaker-corpus normalization. Their experimental results showed that speaker normalization led to the best performance. Zhang et al. [27] also investigated normalization methods and introduced unsupervised learning to handle data sparsity. They proposed that when each corpus is individually normalized, the introduced normalization layers can effectively mitigate the differences among the two datasets. Other publications proposed different kinds of support vector machine (SVM) structures to address the feature distribution mismatch. Hassan et al. [28] proposed modeling the mismatches as a covariate shift. They employed three transfer learning algorithms that apply importance weights (IWs) within an SVM classifier to reduce the effects of covariate shifts. Abdelwahab et al. [29] investigated adaptive and incremental SVMs to reduce the variability in the feature distribution. Their proposed approaches improved the classification performance, even when only a small portion of labeled data was available for adaptation. To generalize the model to unseen languages, Albornoz et al. [30] applied decision-level fusion to improve the recognition accuracy of the SVM classifier. Their system improved the performance of the SER system, even when no data in the target language was available to train the model.

More recently, researchers analyzed deep learning models in cross-corpus SER. In [31], the authors investigated the performance of deep belief networks (DBNs) for cross-corpus SER. They conducted experiments on five emotional datasets and showed that DBNs can learn from many training languages, showing promising performance in SER tasks. Their findings are useful for SER in low-resource languages. In [32], the authors evaluated CNNs and long short-term memory (LSTM) networks using six different speech emotion corpora. Their results indicated that the CNN based model showed better performance on cross-corpus data than the LSTM model. However, since no conclusions can be drawn regarding the extent to which the SER system can generalize across different languages, researchers need to focus on cross-language SER.

2.2 Adversarial Domain Adaptation

To address domain divergence in cross-corpus evaluations, researchers have incorporated domain adaptation methods to transfer emotional information from the source to the target domain representation. In [33], Deng et al. proposed a novel unsupervised domain adaptation method based on adaptive denoising autoencoders for affective speech signal analysis. They trained the denoising autoencoder using unlabeled data from the target domain to learn more robust latent representations. This model effectively and significantly enhanced the emotion classification accuracy in mismatched training and test conditions. In [34], the authors combined a traditional autoencoder with an adversarial autoencoder (AAE) to learn discriminative features from additional data and improved the SER performance with only limited la-

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2023.3290795

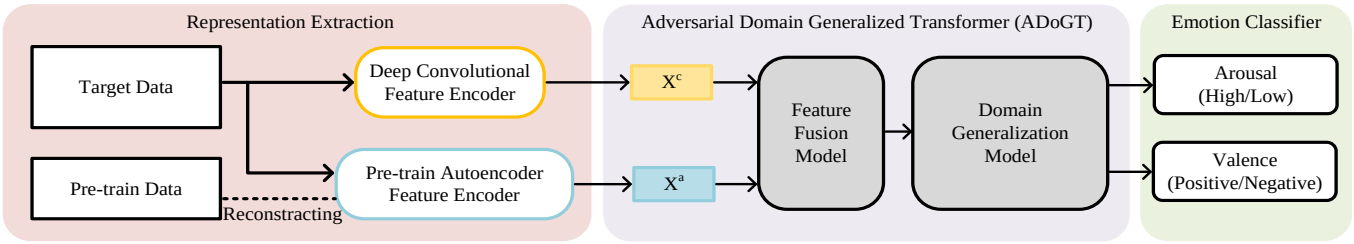


Fig. 1. Flowchart of the proposed Transformer based SER system. For the input utterances, we extract deep representations x^c and x^a from a CNN and a pretrained autoencoder, respectively. We combine the features through the proposed adversarial domain generalized Transformer (ADoGT) and generalize the feature distribution for emotion classification.

beled data. Other researchers proposed eliminating the mismatch between training and testing samples by learning a projection matrix [35]. These works aimed to transform the speech signals in the source and target domains into a similar feature distribution subspace. Gideon et al. [36] proposed adversarial discriminative domain generalization (ADDoG) to learn shared feature representations for the training and test data. They designed a multi-task learning model, which train the model with auxiliary tasks and SER simtaneously. They introduced a critic component as the auxiliary task to encourage the representations of the different datasets to be as close as possible. Their approach outperformed state-of-the-art results in cross-corpus tasks, thus demonstrating the effectiveness of the domain adaptation method.

To achieve domain adaptation and the main classification task simultaneously, Ganin et al. [18] proposed a domain adversarial neural network (DANN) that can be trained using standard backpropagation algorithms. Their proposed structure includes a standard feedforward network and a domain classifier connected to the feature extractor through a gradient reversal layer. This layer reverses the sign of the gradient during backpropagation, thus ensuring that the feature distributions of the two domains are indistinguishable. To build a robust SER system that can not only generalize across speaker information but also other domain information, Abdelwahab et al. [37] proposed applying domain adversarial learning and extracting common representations between the training and test domains. They used domain adversarial learning to extract discriminative feature representations that leveraged unlabeled data in the target dataset and reduced the mismatch between the source and target domains. Their experiments demonstrate that adversarial learning leads to significant improvements in the performance of SER classifiers in which the model is trained with only labeled data from the source domain. This training strategy is aimed at mitigating the influence of nonaffective information. In [38], we incorporated adversarial domain adaptation and eliminated the influence of speaker and corpus information. However, previous works have mainly focused on acted speech. To investigate the performance of our proposed domain generalization method on improvised speech, we include spontaneous data in our model evaluation. In this study, we also present a comprehensive ablation study on DANN subtasks in multi-corpus conditions. Furthermore, we compare the performance of the domain classifiers in the DANN and multi-task learning structure to verify whether domain adversarial learning can make the domain information unlearned to the model.

2.3 Transfer Learning in Affective Computing

Transfer learning is aimed at transferring prior knowledge from different but related source domains to the target domain. Previous publications [39] showed that pretraining representations can effectively improve the robustness and uncertainty of deep learning models. In previous research on affective computing, Ng et al. [40] used a large image dataset to pretrain a CNN based architecture and conducted experiments with two kinds of fine-tuning schemes. The experimental results showed that their model obtained significant improvements over the baseline in facial emotion recognition tasks. Kaya et al. [41] combined the pretrained visual geometry group (VGG) model with a common feature extractor to learn the visual features and then fused these features with the audio features at the decision level to realize multimodal emotion recognition. Their experiments showed that the pretrained model can extract rich features and shows significant improvements over the baseline features. To mitigate the problem of data sparsity in SER, researchers have also investigated several unsupervised transfer learning approaches to transfer prior knowledge from additional datasets. Various publications have shown that autoencoders [42] obtain good performance on image reconstruction tasks and have become widely used in many fields [43], [44]. To extract latent representations for emission recognition, previous studies introduced pretrained autoencoders to extract additional features from unlabeled speech data. These studies evaluated different kinds of autoencoder structures for transferring emotional information from the data utilized for automatic speech recognition [45]. Their models show consistent improvements over baselines with the representations generated by different autoencoder models. In previous works, the features extracted from the pretrained model and feature extractor were usually concatenated to improve the SER performance. In this work, we investigate the impact of a pretrained model in cross-corpus SER and focus on the complementary information of these two kinds of features.

3 Adversarial Domain Generalization

In this section, we describe the overall structure of our approach. As shown in Figure 1, the proposed method is trained in two steps: 1) Representation encoding. We use two branches of feature extractors for feature encoding: a

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2023.3290795

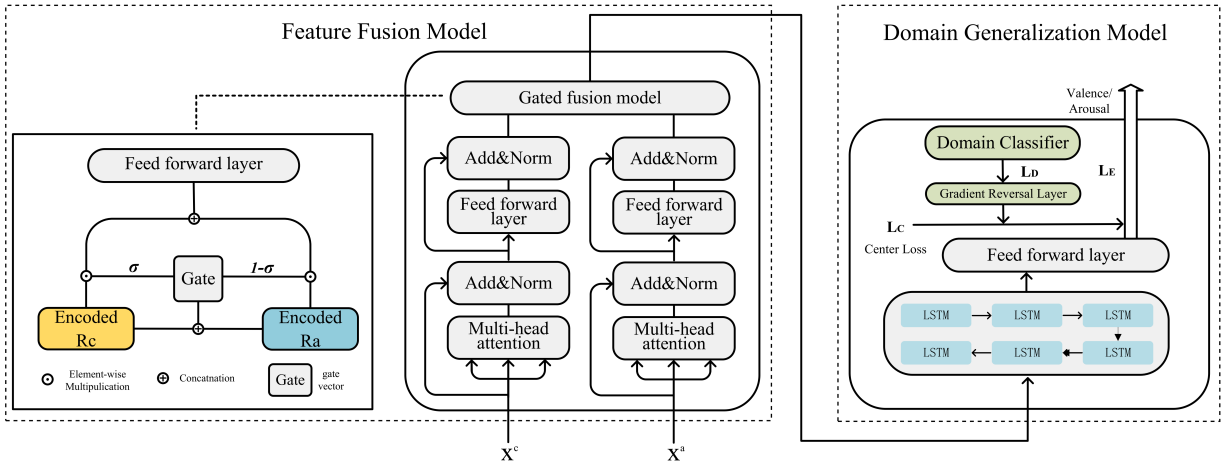


Fig. 2. Our proposed model consists of two parts: (1) Feature fusion. We use a gated fusion model to combine the features learned from the CNN model and pretrained autoencoder. (2) Domain generalization. We modify the emotion classifier as a multi-task DANN network to reduce the domain divergence and combine the center loss with the softmax function for joint supervision. In this Figure, x^c and x^a represent the features learned from the supervised CNN model and unsupervised autoencoder model, respectively. After the Transformer encoder layers, these features are denoted as R^c and R^a, respectively. We use L_E and L_D to represent the emotion and domain classification cross-entropy loss functions, and L_C denotes the center loss function.

supervised CNN architecture and an unsupervised CNN autoencoder. To improve the performance of cross-corpus SER, we combine the output features through our proposed Transformer based gated fusion model. Then, the LSTM layers are used to learn the temporal information. 2) Domain generalization. To reduce the domain divergence of different datasets, we use domain adversarial learning to eliminate the domain information. Furthermore, we incorporate the center loss to obtain more compact intraclass variations for the same emotion. Finally, we use a linear layer with a feature size of 2 for arousal and valence classification.

3.1 Representation Encoding

Learning discriminative features is essential for recognizing emotions. We extract the spectrogram of the emotional utterances as the input to our model. The data preprocessing techniques are described in more detail in Section 4.2.

3.1.1 Supervised Feature Encoder

Previous publications have shown that CNNs can infer hierarchical representations of input utterances that facilitate emotion categorization. As the SER baseline, we use 2D convolutional layers followed by max-pooling layers to learn the spatial information, and then the output features are flattened. For the target data $u = [u_1, u_2, ..., u_n]$, we extract the deep representation $x^c = [x_1^c, x_2^c, ..., x_n^c] \in \mathbb{R}^{d_c \times n}$ from the CNN, and $y^e = [y_1, y_2, ..., y_n] \in \mathbb{R}^{c \times n}$ represents the emotion labels. We extract a d_c-dimensional feature for n utterances, where c is the number of emotions.

3.1.2 Unsupervised Feature Encoder

In this work, we use unsupervised transfer learning to mitigate the problem of data sparsity in SER. Specifically, we incorporate a convolutional autoencoder (AE), which is the most optimized structure for feature modeling, as the pretrained component of the proposed SER system. The AE model is pretrained using unlabeled data u^p, and the objective function is defined as:

$$L_{ae} = \text{argmin}\,||u^p - AE(u^p)||^2 \tag{1}$$

Then, we fine-tune the AE model using unlabeled target data u. The output features of the encoder model are flattened, and we use the latent representation $x^a = [x_1^a, x_2^a, ..., x_n^a] \in \mathbb{R}^{d_a \times n}$ as additional input, where d_a is the dimension of x^a.

3.2 Feature Fusion

As shown in Figure 2, we propose an transformer based gated fusion model to learn the complementary information learned from CNN and autoencoder. Attention mechanism allows a neural network to capture the emotionally salient parts of an input sequence. For x^c and x^a learned from the CNN and autoencoder, we first use two multihead attention branches in the Transformer encoder to reduce the dimension and prevent information loss during feature compression. The attention score is calculated as follows:

$$Q_i = x^{(c,a)} * W_i^Q \tag{2}$$

$$K_i = x^{(c,a)} * W_i^K \tag{3}$$

$$V_i = x^{(c,a)} * W_i^V \tag{4}$$

$$head_i = Softmax(\frac{QK^T}{\sqrt{d}})V \tag{5}$$

where $W_i^Q \in \mathbb{R}^{d_q \times d_x}$, $W_i^K \in \mathbb{R}^{d_q \times d_x}$, and $W_i^V \in \mathbb{R}^{d_q \times d_x}$ are trainable parameter matrices for the attention projections. The final output of the attention layer is generated by concatenating each $head_i$ as follows:

$$x_M^{(c,a)} = Concat(head_1, head_2, ..., head_n) \tag{6}$$

The head number is 8. The outputs of the multihead attention mechanism are x_M^c and x_M^a. We use a fully connected layer to ensure that their dimensions are equal to those of x^a and x^c. To reduce the redundant information of features

learned from the same utterance, we propose a gated fusion model to learn the complementary information of the deep CNN architecture and pretrained model. Our proposed fusion model is described as follows:

$$G = \sigma(W^g[x_M^c, x_M^a] + b^g) \tag{7}$$

$$x = Concat(f(G \odot x_M^c), f((1-G) \odot x_M^a)) \tag{8}$$

where W^g and b^g are the trainable parameters and bias, respectively. σ is the sigmoid activation function, which learns the contributions of the input features, and G denotes the gate vector, which ranges from 0 to 1. In Equation (8), f is the activation function, and $\odot$ represents an elementary product. The gate vector G controls the contributions of the inputs x_M^c and x_M^a by multiplying the corresponding input features and producing filtered representations. Through our modified Transformer based fusion model, we can reduce the irrelevant information in the SER task and learn the emotionally salient parts of the input sequence.

3.3 Domain Generalization

The domain divergence among different datasets has a significantly effect on the feature learning process in the cross-corpus SER system. To improve the SER performance on unseen datasets, we need to generalize the feature distributions of the different datasets. In this work, we introduce domain adversarial learning to eliminate nonaffective information and combine the center loss with an emotion classifier to reduce the intraclass distances of features learned from the same emotion.

3.3.1 Domain Adversarial Learning

In cross-corpus SER, the domain information, including the speaker information, recording conditions, and elicitation method, significantly decreases the recognition performance of deep learning based models. To reduce the domain divergence among different datasets, we incorporate domain adversarial learning to eliminate the nonaffective information. To achieve domain adaptation and feature representation learning with one training process, Ganin et al. introduced a gradient reversal layer (GRL) between the domain classifier and the feature extractor. The GRL multiplies the gradient of the domain classification task by a negative constant γ. In this work, we follow their algorithm and incorporate the supervised emotion classification (L_e) and unsupervised domain classification (L_d) as recognition targets, and unlabeled data in the target corpus are used to train the model. During backpropagation, the domain classifier L_d is trained to make the feature distributions learned from the source and target domains indistinguishable to our model. Through the GRL, we can extract domain invariant representations and thus improve the model generalizability for cross-corpus SER. The overall objective function of our proposed classification model is defined as:

$$L = L_e(x, y^e) + \gamma L_d(x, y^d) \tag{9}$$

where L_e is the loss function of the emotion classifier, which combines the center loss and softmax loss. More details on L_e are provided in Section 3.2.2. Our model can reduce the domain shifts of the feature distributions learned from the source and target datasets with unsupervised domain adaptation for the source and target data. In this specific task, by incorporating this training strategy with our supervised feature extraction model, the domain-invariant features can retain discriminative information for emotion classification. The loss function of the domain classifier is defined as:

$$L_d = L_{d_1}(x, y^{d_1}) + L_{d_2}(x, y^{d_2}) + ... + L_{d_n}(x, y^{d_n}) \tag{10}$$

where n is the number of domain classifiers and $y^{d_i} = [y_1, y_2, ..., y_n] \in \mathbb{R}^{c^{d_i} \times n}$ represents the corresponding labels. We explore different DANN subtasks to determine the optimal model structure for SER. By identifying a saddle point that minimizes L_e and maximizes L_d, our proposed feature extractor can reduce the domain divergence and learn better convergent features.

3.3.2 Center Loss

The center loss is combined with the emotion classifier to reduce the intraclass distance, and we incorporate the softmax loss and center loss as joint supervision for the emotion classifier L_e. The softmax loss function is commonly used in SER systems for identifying decision boundaries between different emotions [46]. In this study, although we define the same emotion annotations for the training and test samples, the feature distributions of different datasets are difficult to separate. This situation makes cross-corpus SER more challenging than common closed-set identification tasks. To mitigate this problem, we introduce the center loss to learn the class center c for each emotion category, thus reducing the intraclass distances in the feature distribution. This loss function is calculated as the Euclidean distance between the input feature and the corresponding class center.

$$Center(x, c) = \frac{1}{M}\sum_{i=1}^{N} ||x^{(i,y_j)} - c^{y_j}||^2 \tag{11}$$

$$\Delta c^{y_j} = \frac{\sum_{i=1}^{M^{y_j}} (c^{y_j} - x^{(i,y_j)})}{1+m} \tag{12}$$

where M and M^{y_j} are the total number of mini-batches and the j_{th} emotion category in the batch, respectively. N is the number of emotion classes. The new class center is updated by Δc^{y_j}, which is trained for every mini-batch. The overall objective function of the emotion classifier is defined as:

$$L_E(x, y^e) = \lambda Softmax(x, y^e) + (1-\lambda)Center(x, c) \tag{13}$$

We set λ to 0.7 to control the weight of each loss term. By combining the center loss with the softmax loss to jointly optimize our model, we can extract more robust feature representations that generalize across datasets.

4 Experimental Setup

Four emotional datasets are used to evaluate the generalizability of our model: IEMOCAP, MSP-IMPROV, EMODB, and FAU-AIBO. All the datasets are publicly available. The datasets cover different languages and elicitation methods and are thus valuable for evaluating our model. We first present the main attributes of each dataset and the emotion labels in this study (Table 1). Then, we introduce two well-known unlabeled datasets, which are used to pretrain the autoencoder. This section also includes the data preprocessing techniques and model configuration.

TABLE 1
Overview of the four emotion corpora. For IEMOCAP, the elicitation type contains both acted and natural, and the lexical content contains both scripted and improvised.

Corpus	Language	#m #f	Rate	Type	Content	Total	Valence		Arousal	
							Negative	Positive	Low	High
IEMOCAP	English	5 5	16 kHz	Hybrid	Hybrid	5531	3344	2187	2792	2739
MSP-Improv	English	6 6	44.1 kHz	Acted	Fixed	8438	4546	3892	3660	4778
Emo-DB	German	5 5	16 kHz	Acted	Fixed	535	385	150	268	267
Fau-aibo	German	30 21	16 kHz	Natural	Spontaneous	18216	5093	13123	15835	2318

TABLE 2
Emotion mapping from discrete labels to binary arousal

Datasets	Low	High
IEMOCAP	Neutral, Sad	Angry, Happy
MSP-IMPROV	Neutral, Sad	Angry, Happy
EMODB	Bordorm, Disgust, Neutral, Sad	Angry, Happy, Fear
FAU-AIBO	Neutral, Rest, Emphatic	Angry, Joy

TABLE 3
Emotion mapping from discrete labels to binary valence

Datasets	Negative	Positive
IEMOCAP	Angry,Sad	Happy,Neutral
MSP-IMPROV	Angry,Sad	Happy,Neutral
EMODB	Angry,Bordorm, Disgust,Fear,Sad	Happy,Neutral
FAU-AIBO	Angry,Emphatic	Neutral,Joy,Rest

4.1 Datasets

The **IEMOCAP** dataset: The Interactive Emotional Dyadic Motion Capture database [47] contains 12 hours of audio-visual data, including audio, video, and facial motion information, and textual transcriptions from 10 speakers. The audio was recorded using two high-quality microphones with a 48 kHz sampling rate and then downsampled to 16 kHz. In each session, one male and one female performed a series of scripts or improvisational scenarios. For each speech utterance, three annotators assigned the categorical labels. We used 5531 utterances from the scripted and improvised audio data for our experiments. We implemented the common practice of merging "happy" and "excited" into one emotion class "happy" [45], [48], [49]; thus, the emotion labels in this dataset are happy, sad, angry, and neutral.

The **MSP-IMPROV** dataset: MSP-IMPROV [50] is a multimodal emotional database that includes recordings of actors interacting in dyadic sessions. The actors aim to control the lexical content of each sentence while displaying natural emotional expressions. The corpus consists of 8,438 utterances (8.9 hours) of emotional sentences recorded from 12 actors. The audio data of each actor was recorded with a collar microphone with a 48 kHz sampling rate and then downsampled to 44.1 kHz. All the audio data were grouped into six sessions, and each session has one male and one female actor. The categorical labels were collected using crowdsourcing on Amazon Mechanical Turk. The emotion categories in this dataset are happy, sad, angry, and neutral.

The **MSP-IMPROV** dataset: MSP-IMPROV [50] is a multimodal emotional database recorded from actors interacting in dyadic sessions. The author aims to control the lexical content of each sentence while promoting the naturalness of emotion expression. The corpus consists of 8,438 utterances (8.9 hours) of emotional sentences recorded from 12 actors. Each actor used a collar microphone to record speech at 48 kHz and then downsampled it to 44.1 kHz. All the audio data are grouped into six sessions and each session has one male and one female actor. The categorical labels are collected using crowdsourcing on Amazon Mechanical Turk. The emotion categories in this dataset are also happy, sad, angry, and neutral.

The **EMODB** dataset: The Berlin Emotional Speech database [51] includes data from ten professional actors obtained in a recording environment. The spoken content includes 10 predefined emotionally neutral sentences in German, and the actors were asked to express each sentence in seven emotional states (neutral, boredom, disgust, sadness, anger, happiness, and fear). The categorical labels were collected according to the intended emotional state. This corpus contains a total of 535 utterances, which had an agreement rate higher than 84.3% in a listening experiment with 20 participants (10 male and 10 female).

The **FAU-AIBO** dataset: The FAU Aibo Emotion Corpus [52] was recorded to collect spontaneous audio data with sufficient emotion expression. This dataset contains spontaneous recordings of 51 children interacting with the Sony robot Aibo. Thirty female and 21 male pupils were instructed to talk with Aibo, and then five experts annotated the recorded speech according to predefined emotion categories. We choose 18216 utterances used for the Interspeech Emotion Challenge, including five emotion categories (angry, emphatic, neutral, joyful, and rest).

The annotated labels of these datasets are inconsistent in this study, and we chose arousal and valence to generate more interpretable emotion classification categories. Although IEMOCAP and MSP-IMPROV have continuous labels for arousal and valence, to maintain consistency with other datasets, we followed Schuller et al. [12] and mapped the discrete emotion labels to binary arousal and valence (Tables 2 and 3). In this work, we choose Librispeech [53] and MUST-C DE to pretrain the autoencoder [54]. Lib-

rispeech is commonly used in speaker identification and automatic speech recognition tasks; it contains 1000 hours of English speech read from audiobooks. We selected the 360-hour subset with high-quality recording conditions. MUST-C consists of audio, transcriptions and translations of English TED talks. We used the MUST-C DE subset, which contains 408 hours of German translations from English TED Talks.

4.2 Model Configuration

During data preprocessing, to match the sampling rate, all the datasets are downsampled to 16 kHz. We use a 256-length Hamming window with 128 overlaps to calculate the audio spectrogram of the input features using MATLAB. For the variable-length inputs, we define the maximum length of the time dimension as 700. Spectrograms with shorter lengths are padded with zeros to the fixed length, and the redundant parts are masked during training. After a short Fourier transform, the time $\times$ frequency of the input spectrogram is calculated as 700 $\times$ 129. Our experiments are implemented using PyTorch [55]. To ensure consistency with the baseline, we use three CNN layers followed by max-pooling as the supervised feature extractor in all comparative experiments. In experiments with the pretrained model, we use five convolutional layers as the encoder and the corresponding five deconvolutional layers as the decoder. Moreover, we use the output of the encoder as the latent representation. The learned features of the supervised and unsupervised feature extractors are then flattened and fed into the Transformer encoder layers and gated fusion model. Next, we use two bidirectional LSTM layers with 128 units to learn the sequence information, and a dropout layer with a 0.5 dropout rate is used to prevent overfitting. The LSTM output is fed into a fully connected layer with a softmax function for classification. We employed Adadelta as the optimizer, and the mini-batch size was set to 128. To ensure that SER was the training objective of the total loss function, the weight parameter of each domain classifier ranged from 0 to 0.5. For the pretrained model, we pretrain the autoencoder model and then fine-tune this model using unlabeled data from the source domain. During feature extraction, we maintain a fixed weight and bias and use the target data to extract the bottleneck features.

For multi-corpus experiments, all four datasets are combined. We split the data into a training set (80%) and a test set (20%). The models are evaluated using test data from each corpus. Note that there is no speaker overlap between the training and testing data. For cross-corpus experiments, no labeled data from the target corpus are used for training.

5 Experiments and Evaluations

In this study, we design several experiments to evaluate our proposed approach. First, we investigate the effect of different domain adaptation subtasks in Section 5.1. Then, we compare the proposed model with the CNN-LSTM baseline in Section 5.2. To determine the impact of each component on the overall system, ablation studies for the proposed domain generalization method and Transformer based gated fusion network are presented in Sections 5.3 and 5.4. Finally, we compare the performance of the domain classifier in the DANN and multi-task learning model in Section 5.5. In this study, we use the unweighted accuracy (UA) as evaluation measure, which can avoid the influence of data imbalence in each emotion.

TABLE 4
Multi-corpus evaluation results for learning the impact of domain adaptation methods. In DANN, we choose the domain classifier with best performance. In DG, we combine center loss with softmax loss function as joint supervision.

	Arousal			Valence		
Model	CNN	DANN	DG	CNN	DANN	DG
IEM	73.28	**78.56**	78.35	70.38	73.09	**75.19**
MSP	60.80	64.37	**65.71**	60.46	62.58	**62.70**
EMO	90.21	**93.74**	92.58	62.74	65.73	**67.24**
FAU	53.77	**55.81**	55.21	60.31	**63.74**	61.62
Avg.	69.51	**73.12**	72.96	63.47	66.29	**66.84**

5.1 Experiment 1: Multi-corpus Evaluation

As introduced in Section 4.2, all four datasets are used in this multi-corpus evaluation. In this section, only the deep CNN model is used as the feature extractor. We also evaluate the performance of the DANN and proposed domain generalization (DG) method, which combines the the center loss and softmax as joint supervision for SER. The DANN and CNN are compared in Table 4, and the results show that the domain adversarial learning method learns more discriminative features in both arousal and valence recognition. Thus, incorporating the center loss and DANN for domain generalization leads to promising performance with these four datasets, especially in valence recognition.

To better understand the domain adversarial learning method, we design four types of DANN subtasks. We hypothesize that different domain recognition targets can benefit the SER system when the corresponding factors lead to domain divergence in the feature extraction process. We conduct a multi-corpus evaluation to explore the effectiveness of different domain classifiers on specific datasets. As the training data contain domain information from all four datasets, the results can intuitively reflect the effect of each DANN structure in certain testing conditions.

We present seven experiments to explore the impact of the speaker, gender, language, and elicitation type. There are two kinds of structures in this experiment: a DANN with one domain classifier branch (D_1) and a DANN with two domain classifier branches (D_2). Previous studies on SER have demonstrated that the speaker information has a significant influence on the classification results. If the speakers in the training and test data overlap, the SER model shows better performance than a model trained with a speaker-independent validation strategy. Therefore, the first subexperiment with D_1 uses the speaker as the domain classifier. Since the gender and language information can reflect the speaker information, in D_1, we also train domain classifiers for language and gender classification. In addition to the speaker information, the elicitation strategy (acted or spontaneous) has a great impact on the performance of SER. Therefore, an elicitation type classifier (type) is

TABLE 5
Multi-corpus evaluation results for learning the impact of different domain classifiers. In this table, D1 has one branch of the domain classifier. In D2 models, we add type as an additional domain learning target to D1.

Test	Arousal							Valence						
	type	gender		language		speaker		type	gender		language		speaker	
		D1	D2	D1	D2	D1	D2		D1	D2	D1	D2	D1	D2
IEM	76.23	**78.56**	77.85	76.45	76.3	73.62	75.28	**73.09**	71.63	70.86	71.86	72.76	70.53	71.25
MSP	63.09	61.84	**64.37**	62.13	63.19	61.54	62.93	61.53	**62.58**	61.35	61.11	61.97	58.18	60.13
EMO	**93.74**	91.58	91.35	91.2	92.52	88.77	91.26	64.16	62.41	63.42	63.92	**65.73**	61.35	62.27
FAU	53.31	53.64	**55.81**	54.64	55.26	53.8	52.95	62.38	62.07	62.37	62.15	**63.74**	61.28	60.74
Avg.	71.59	71.66	**72.35**	71.11	71.82	69.43	70.61	65.29	64.67	64.50	64.76	**66.05**	62.84	63.60

also included. In D_2, we combine the speaker, gender, and language classifiers with the type classifier.

As shown in Table 5, for both arousal and valence recognition, the language (D2) classifiers show the highest average recognition performance. To reflect the advantages of domain adaptation methods, we use language and type classifiers in the DANN model in the following experiments. For each dataset, the best DANN structure includes type, gender (D2) and language (D2). These results indicate the effectiveness of domain adversarial learning based on the elicitation type. In previous studies, speaker and gender classification were commonly employed in multi-task learning models. However, these two classifiers cannot realize the best performance in domain adaptation methods. This finding may indicate that eliminating the corresponding information does not benefit SER system.

5.2 Experiment 2: Cross-corpus Evaluation

Then, we utilize the following experiments to investigate the effectiveness of the proposed approach in cross-corpus evaluation. CNN-LSTM models have been utilized in many previous SER publications, and we choose this model as the baseline. Then, we compare the performance of the proposed ADoGT approach with that of the baseline system. Our proposed DG model can generate a common feature subspace for different domains by combining the DANN and center loss. Furthermore, we use a pretrained autoencoder as an additional branch in our proposed model and apply the proposed Transformer based model for feature fusion. For each model, we provide 13 experimental results for evaluation (training with one dataset and testing with the other datasets and the average performance).

Table 6 shows the overall cross-corpus evaluation results. The results suggest that both models realize better performance on arousal recognition tasks, especially in cross-lingual experiments (IEMOCAP and EMODB). These results indicate that arousal information is easier to learn for deep learning models than valence information, which is consistent with [27], [56]. Compared with the baseline system, the experimental results demonstrate the advantages of the proposed generalization method and pretrained model in cross-corpus SER. We observe that for speech recorded in laboratory environments, the proposed Transformer based domain generalization model significantly improved recognition performance (e.g. more than 6% for both arousal and valence recognition when train on IEMOCAP and test on EMODB). However, due to the poor performance of both approaches in certain experiments (e.g., training on FAU-AIBO and testing on other datasets, valence recognition between MSP-IMPROV and EMODB), the improvement in the average performance is not significant. In the following experiments, we investigate the impact of each component in the proposed approach.

TABLE 6
Cross-corpus evaluation results for analysing the effectiveness of proposed domain generalization method

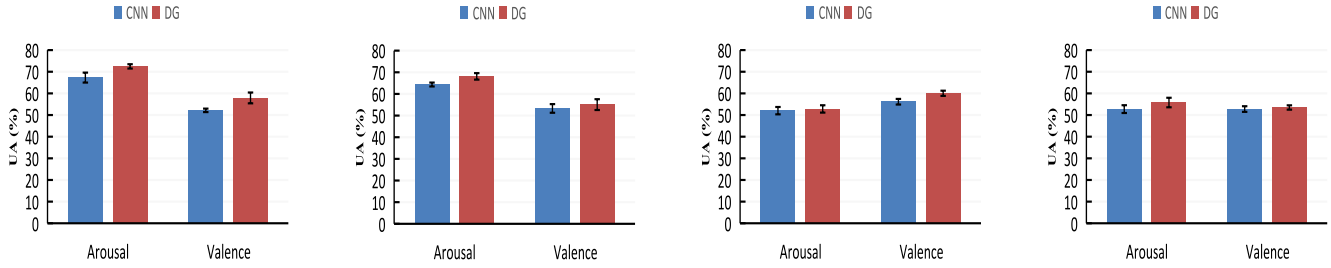
train on	test on	**Arousal**		**Valence**	
		CNN	DoGAT	CNN	DoGAT
IEM	MSP	57.29	61.83	55.73	59.42
	EMO	67.3	73.53	52.48	58.73
	FAU	52.02	52.06	56.15	60.41
MSP	IEM	61.78	63.75	56.81	57.12
	EMO	55.37	59.82	54.52	58.15
	FAU	54.88	55.67	52.01	55.87
EMO	IEM	64.35	68.23	53.30	58.07
	MSP	52.33	55.49	50.32	50.37
	FAU	51.62	51.58	54.04	59.49
FAU	IEM	52.74	55.29	52.77	53.02
	MSP	52.61	57.38	50.28	50.00
	EMO	54.69	56.75	57.46	59.46
Model Avg.		56.42	59.28	53.82	56.68

5.3 Experiment 3: Impact of Domain Generalization

We hypothesize that the domain generalization method effectively reduces the domain divergence among different datasets. To compare the proposed domain generalization model with the baseline system, we conduct cross-corpus experiments using IEMOCAP with MSP-IMPROV and EMODB. For both models, we use only a supervised feature extraction model with no additional feature inputs and apply consistent CNN and LSTM hyperparameters. We repeat this experiment five times and report the mean and standard deviation.

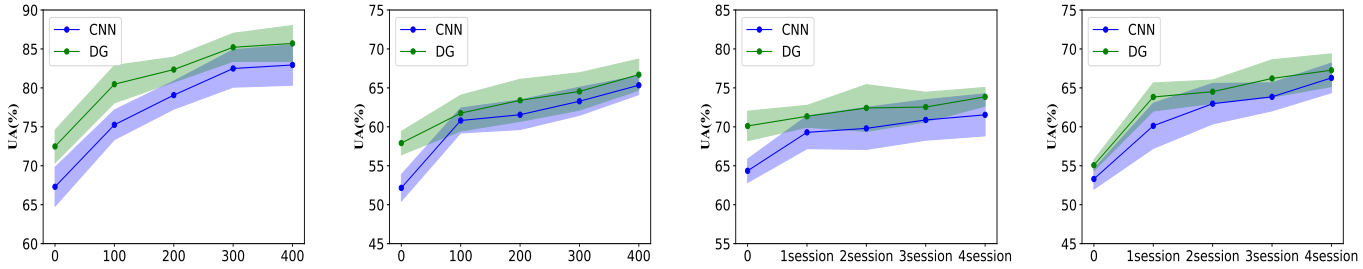
The results of the comparisons are presented in Figure 3. We use gender and type and language and type as domain classifiers for monolingual and cross-lingual experiments, respectively. We observe that for both arousal and valence classification, our proposed domain generalization method outperforms the single-task learning baseline. This finding

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2023.3290795



(a) IEMOCAP to EMODB (b) EMODB to IEMOCAP (c) IEMOCAP to FAU-AIBO (d) FAU-AIBO to IEMOCAP

Fig. 3. Impact of proposed domain generalization model. We use elicitation type and language as domain classifiers for IEMOCAP and FAU-AIBO experiments, IEMOCAP and EMODB experiments, respectively.



(a) IEMOCAP to EMODB (a) (b) IEMOCAP to EMODB (v) (c) EMODB to IEMOCAP (a) (d) EMODB to IEMOCAP (v)

Fig. 4. Cross-corpus experimental results of IEMOCAP and EMODB. This experiment includes an increasing number of labelled data from the target dataset. In (a) and (c), a is the arousal recognition results. In (b) and (d), v is the valence recognition results.

indicates that the domain adaptation task can be adopted for the SER system. Moreover, introducing the center loss effectively improves the generalizability of the feature representation. Interestingly, we observe that when the model is trained with IEMOCAP and tested with FAU-AIBO, the valence recognition performance is better than the arousal recognition performance. We assume that the main reasons for this result are that this dataset was recorded in daily environments and the speakers did not realize that they were recording emotional speech. Therefore, the activation states of most utterances are lower than those of the utterances in the other corpora. Furthermore, both models achieve better recognition performance in the cross-corpus experiments with IEMOCAP and EMODB than IEMOCAP and FAU-AIBO. This result potentially indicates the difficulty of SER with improvised (spontaneous) data. In this experiment, we also incorporate some of the labeled data in the target dataset to further evaluate our model. As depicted in Figure 4, our model can use the target data to learn shared feature representations for different domains and achieve better performance. This Figure also demonstrates the effectiveness of our model in multi-corpus evaluation, where most of the labeled data in the target corpus are used for training.

5.4 Experiment 4: Impact of Feature Fusion Model

The Transformer based gated fusion model can prevent information loss during dimension reduction and learn the complementary information of the input features. In this experiment, we use IEMOCAP and MSP-IMPROV to evaluate the effectiveness of the proposed method. We present both within-corpus experiments (80% of the data in IEMOCAP for training and 20% for testing) and cross-corpus experiments (training with IEMOCAP and testing with MSP-IMPROV). The input model includes a supervised CNN/DANN and a pretrained AE for additional feature extraction. Note that only softmax loss function for emotion classification is used, and domain generalization methods is not included in this experiment. We compare the concatenation method and Transformer based feature fusion model, and the results are presented in Table 7.

TABLE 7
The comparasion of different feature fusion methods. The model is trained on IEMOCAP, we present the within-corpus results (left hand) and cross-corpus results (right hand).

Model	IEMOCAP		MSP	
	Arousal	Valence	Arousal	Valence
CNN	75.66	70.58	57.29	55.73
Concatenate	77.86	72.43	58.27	56.24
Transformer_C	78.35	74.25	59.05	57.83
TGFM	**79.60**	**74.73**	**60.46**	**58.81**

A comparison of the concatenation based model and CNN baseline results shows that using a large amount of unlabeled data to pretrain the autoencoder can improve the

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2023.3290795

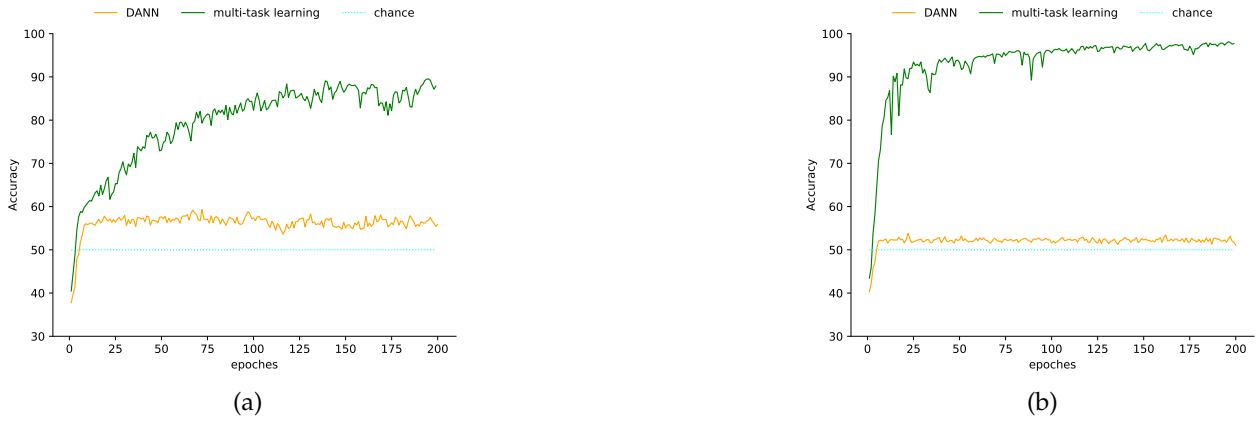


Fig. 5. Comparison of the domain classifier in the multi-task learning model and domain adversarial neural network with the same model structure. We define the weight parameter of the auxiliary task as 0.01 in (a) and 0.1 in (b).

performance of SER, which is consistent with the conclusions of previous studies [23], [57]. In both the within-corpus and cross-corpus evaluations, our proposed Transformer based gated fusion model (TGFM) outperforms the concatenation based method by more than 1.74%. Compared with the CNN baseline, we observe that including the Transformer model improves the results of the model with the concatenation layer for feature fusion (Transformer_C). This finding demonstrates the discriminability of the attention mechanism for learning emotional information.

5.5 Experiment 5: Comparasion of Domain Classifier

To elucidate the effects of domain adversarial learning, we focus on the performance of the domain classifier in the DANN and multi-task learning model to investigate the effect of the gradient reversal layer. In previous studies, researchers incorporated the recognition of other speaker attributes to obtain rich transcriptions, and their experiments demonstrated the effectiveness of this training scheme. Among all paralinguistic information and emotion attributes, the impact of gender recognition has been analyzed most often in the literature [58], [59]. In this experiment, both models are trained with IEMOCAP and tested with MSP-IMPROV, and we choose gender classification as the domain classifier. To understand the effect of the GRL, we record the model accuracy with the training set, and the weight of the domain classifier is defined as 0.01 and 0.1.

As depicted in Figure 5 (a) and (b), in the multi-task learning model, the recognition performance of the gender classifier is significantly higher than that of the DANN model. This performance gap verifies the effectiveness of domain adversarial learning in reducing the domain divergence among different datasets. This experiment also suggests that when the weight parameter of the domain classifier is greater than 0.1, domain classification maintains the chance level and thus makes the features from different domains indistinguishable.

6 Conclusion and Future Work

This study addresses domain divergence in cross-corpus SER by incorporating domain adversarial learning to feature extraction model and jointly training the emotion classifier with center loss and softmax loss function. This study introduced an novel Transformer based gated fusion model to retain emotional information during feature compression and learns the contributions of features learned from the pretrained model and supervised feature extractor. To meet the need in real-world scenarios, this study evaluated the proposed adversarial domain generalized Transformer (ADoGT) in two languages (English and German) and two elicitation types (spontaneous and acted). To verify the impact of domain adversarial learning, this study provide the comparison of domain classifier in domain adversarial neural network (DANN) and multi-task learning.

Experimental results demonstrate that our proposed model improves the average recognition performance by 2.86% in the cross-corpus condition. To learn the influence of each domain factor on the SER, we present the results of multi-corpus experiments using the DANN with different domain recognition targets. Compared with the baseline model, the proposed domain generalization model obtains better recognition performance by reducing the influence of domain divergence. Moreover, ablation studies show the effectiveness of the Transformer based gated fusion model in feature-level fusion tasks. Compared with the concatenation based model, our approach utilizes the complementary information in features learned from the same utterance and thus prevents information loss.

This study mainly focuses on DANN and the center loss to reduce the domain divergence and address interdomain variations. Both training methods aim to learn discriminative features for SER and make the nonaffective information indistinguishable to the model. Other publications have shown thatmulti-task learning can benefit SER tasks by achieving rich transcriptions. These studies define the attribute factor as a subtask and share the information across tasks to promote the SER. Combining DANN and multi-task learning can potentially provide further insight into how other information influences the SER performance in certain scenarios. Given the many learning factors, this topic requires continuous attention from researchers in affective computing. In the future, we plan to investigate the optimal auxiliary recognition target of these two approaches.

REFERENCES

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal processing magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[2] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: a review," *International journal of speech technology*, vol. 15, no. 2, pp. 99–117, 2012.

[3] S. Ramakrishnan and I. M. El Emary, "Speech emotion recognition approaches in human computer interaction," *Telecommunication Systems*, vol. 52, no. 3, pp. 1467–1478, 2013.

[4] R. Feldman, "Techniques and applications for sentiment analysis," *Communications of the ACM*, vol. 56, no. 4, pp. 82–89, 2013.

[5] M. Wimmer, B. Schuller, D. Arsic, B. Radig, and G. Rigoll, "Low-level fusion of audio and video feature for multi-modal emotion recognition," in *Proc. 3rd Int. Conf. on Computer Vision Theory and Applications VISAPP, Funchal, Madeira, Portugal*, 2008, pp. 145–151.

[6] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech emotion recognition using deep learning techniques: A review," *IEEE Access*, vol. 7, pp. 117 327–117 345, 2019.

[7] J. Zhang, Z. Yin, P. Chen, and S. Nichele, "Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review," *Information Fusion*, vol. 59, pp. 103–126, 2020.

[8] B. J. Abbaschian, D. Sierra-Sosa, and A. Elmaghraby, "Deep learning techniques for speech emotion recognition, from databases to models," *Sensors*, vol. 21, no. 4, p. 1249, 2021.

[9] Y. Zhang, J. Du, Z. Wang, J. Zhang, and Y. Tu, "Attention based fully convolutional network for speech emotion recognition," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2018, pp. 1771–1775.

[10] M. Abdelwahab and C. Busso, "Study of dense network approaches for speech emotion recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5084–5088.

[11] B. Liu, "Sentiment analysis and opinion mining," *Synthesis lectures on human language technologies*, vol. 5, no. 1, pp. 1–167, 2012.

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

[13] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7167–7176.

[14] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for speech emotion recognition," *Neural Networks*, vol. 92, pp. 60–68, 2017.

[15] D. Nguyen, K. Nguyen, S. Sridharan, D. Dean, and C. Fookes, "Deep spatio-temporal feature fusion with compact bilinear pooling for multimodal emotion recognition," *Computer Vision and Image Understanding*, vol. 174, pp. 33–42, 2018.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[17] S. Latif, J. Qadir, and M. Bilal, "Unsupervised adversarial domain adaptation for cross-lingual speech emotion recognition," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2019, pp. 732–737.

[18] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The journal of machine learning research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[19] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, p. e1253, 2018.

[20] A. Yadav and D. K. Vishwakarma, "Sentiment analysis using deep learning architectures: a review," *Artificial Intelligence Review*, vol. 53, no. 6, pp. 4335–4385, 2020.

[21] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *European conference on computer vision*. Springer, 2016, pp. 499–515.

[22] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, 2018.

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

[24] S. M. Feraru, D. Schuller *et al.*, "Cross-language acoustic emotion recognition: An overview and some tendencies," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2015, pp. 125–131.

[25] R. Milner, M. A. Jalal, R. W. Ng, and T. Hain, "A cross-corpus study on speech emotion recognition," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 304–311.

[26] M. Gerczuk, S. Amiriparian, S. Ottl, and B. W. Schuller, "Emonet: A transfer learning framework for multi-corpus speech emotion recognition," *IEEE Transactions on Affective Computing*, 2021.

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

[28] A. Hassan, R. Damper, and M. Niranjan, "On acoustic emotion recognition: compensating for covariate shift," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1458–1468, 2013.

[29] M. Abdelwahab and C. Busso, "Supervised domain adaptation for emotion recognition from speech," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5058–5062.

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

[31] S. Latif, R. Rana, S. Younis, J. Qadir, and J. Epps, "Transfer learning for improving speech emotion classification accuracy," *arXiv preprint arXiv:1801.06353*, 2018.

[32] J. Parry, D. Palaz, G. Clarke, P. Lecomte, R. Mead, M. Berger, and G. Hofer, "Analysis of deep learning architectures for cross-corpus speech emotion recognition." in *INTERSPEECH*, 2019, pp. 1656–1660.

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

[34] 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 Transactions on Affective Computing*, 2020.

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

[36] J. Gideon, M. G. McInnis, and E. M. Provost, "Improving cross-corpus speech emotion recognition with adversarial discriminative domain generalization (addog)," *IEEE Transactions on Affective Computing*, vol. 12, no. 4, pp. 1055–1068, 2019.

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

[38] Y. Gao, S. Okada, L. Wang, J. Liu, and J. Dang, "Domain-invariant feature learning for cross corpus speech emotion recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6427–6431.

[39] Y. Bengio, "Deep learning of representations for unsupervised and transfer learning," in *Proceedings of ICML workshop on unsupervised and transfer learning*. JMLR Workshop and Conference Proceedings, 2012, pp. 17–36.

[40] H.-W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler, "Deep learning for emotion recognition on small datasets using transfer learning," in *Proceedings of the 2015 ACM on international conference on multimodal interaction*, 2015, pp. 443–449.

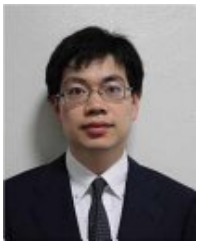
[41] H. Kaya, F. Gürpınar, and A. A. Salah, "Video-based emotion recognition in the wild using deep transfer learning and score fusion," *Image and Vision Computing*, vol. 65, pp. 66–75, 2017.

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

[43] L. Deng, M. L. Seltzer, D. Yu, A. Acero, A.-r. Mohamed, and G. Hinton, "Binary coding of speech spectrograms using a deep

auto-encoder," in *Eleventh Annual Conference of the International Speech Communication Association*. Citeseer, 2010.
[44] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder." in *Interspeech*, vol. 2013, 2013, pp. 436–440.
[45] S. Latif, R. Rana, J. Qadir, and J. Epps, "Variational autoencoders for learning latent representations of speech emotion: A preliminary study," *arXiv preprint arXiv:1712.08708*, 2017.
[46] H. Meng, T. Yan, F. Yuan, and H. Wei, "Speech emotion recognition from 3d log-mel spectrograms with deep learning network," *IEEE access*, vol. 7, pp. 125 868–125 881, 2019.
[47] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.
[48] V. Rozgić, S. Ananthakrishnan, S. Saleem, R. Kumar, and R. Prasad, "Ensemble of svm trees for multimodal emotion recognition," in *Proceedings of the 2012 Asia Pacific signal and information processing association annual summit and conference*. IEEE, 2012, pp. 1–4.
[49] P. Tzirakis, J. Zhang, and B. W. Schuller, "End-to-end speech emotion recognition using deep neural networks," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5089–5093.
[50] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. M. Provost, "Msp-improv: An acted corpus of dyadic interactions to study emotion perception," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, 2016.
[51] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, B. Weiss *et al.*, "A database of german emotional speech." in *Interspeech*, vol. 5, 2005, pp. 1517–1520.
[52] S. Steidl, *Automatic classification of emotion related user states in spontaneous children's speech*. Logos-Verlag Berlin, Germany, 2009.
[53] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.
[54] R. Cattoni, M. A. Di Gangi, L. Bentivogli, M. Negri, and M. Turchi, "Must-c: A multilingual corpus for end-to-end speech translation," *Computer Speech & Language*, vol. 66, p. 101155, 2021.
[55] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.
[56] S. Zhang, X. Zhao, and Q. Tian, "Spontaneous speech emotion recognition using multiscale deep convolutional lstm," *IEEE Transactions on Affective Computing*, 2019.
[57] S. E. Eskimez, Z. Duan, and W. Heinzelman, "Unsupervised learning approach to feature analysis for automatic speech emotion recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5099–5103.
[58] Z.-Q. Wang and I. Tashev, "Learning utterance-level representations for speech emotion and age/gender recognition using deep neural networks," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 5150–5154.
[59] A. Nediyanchath, P. Paramasivam, and P. Yenigalla, "Multi-head attention for speech emotion recognition with auxiliary learning of gender recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7179–7183.

Yuan Gao is currently pursuing a Ph.D. degree at the Department of Intelligence Science and Technology, Kyoto University, Kyoto, Japan. He received an M.S. degree from both Tianjin University, Tianjin, China, and the Japan Advanced Institute of Science and Technology, Ishikawa, Japan, in 2022. His research interests include speech signal processing and multimodal emotion recognition.



Longbiao Wang received his Dr.Eng. degree from Toyohashi University of Technology, Japan, in 2008. He was an Assistant Professor in the faculty of Engineering at Shizuoka University, Japan, from April 2008 to September 2012. He was an Associate Professor at Nagaoka University of Technology, Japan from Oct. 2012 to Aug. 2016. He is currently a Professor, Director of Tianjin Key Laboratory of Cognitive Computing and Application and vice Dean of School of Artificial Intelligence at Tianjin University, China. His research interests include robust speech recognition, speaker recognition, acoustic signal processing and affective computing. He is a member of IEEE.

Jiaxing Liu He is currently pursuing the PhD degree with the College of Intelligence and Computing, Tianjin University, Tianjin, China. Now is the an exchanging student in Nanyang Technological University. His research interests include speech emotion recognition, multimodal emotion recognition, and natural language processing (sentiment analysis).

Jianwu Dang Prof. Jianwu Dang graduated from Tsinghua Univ., China, in 1982, and got his M.S. degree at the same university in 1984. He worked for Tianjin Univ. as a lecture from 1984 to 1988. He was awarded the PhD degree from Shizuoka Univ., Japan in 1992. He worked for ATR Human Information Processing Labs., Japan, as a senior researcher from 1992 to 2001. He joined the University of Waterloo, Canada, as a visiting scholar for one year from 1998. He worked for Japan Advanced Institute of Science and Technology (JAIST) as a professor from 2001-2022. He joined the Institute of Communication Parlee (ICP), Center of National Research Scientific, France, as a research scientist the first class from 2002 to 2003. Since 2009, he has joined Tianjin University, Tianjin, China. His research interests are in all the fields of speech science including speech signal processing, disorder speech and speech cognitive functions.

Shogo Okada directs the Social Signal and Interaction Group at the Japan Advanced Institute of Science and Technology (JAIST) in Japan and is an associate professor at JAIST. He obtained his Ph.D. in 2008 from Tokyo Institute of Technology in Japan. In 2008 and 2011, he joined the Kyoto University, Tokyo Institute of Technology, as an assistant professor. He joined IDIAP Research Institute in Switzerland as a visiting faculty member in 2014. His research interests include social signal processing, human dynamics, multimodal interaction, and machine learning. He is a member of IEEE and ACM .