frontiers in Computer Science

REVIEW
published: 28 February 2020
doi: 10.3389/fcomp.2020.00009

A Review of Generalizable Transfer Learning in Automatic Emotion Recognition

Kexin Feng and Theodora Chaspari*

HUman Bio-Behavioral Signals (HUBBS) Lab, Texas A&M University, College Station, TX, United States

Automatic emotion recognition is the process of identifying human emotion from signals such as facial expression, speech, and text. Collecting and labeling such signals is often tedious and many times requires expert knowledge. An effective way to address challenges related to the scarcity of data and lack of human labels, is transfer learning. In this manuscript, we will describe fundamental concepts in the field of transfer learning and review work which has successfully applied transfer learning for automatic emotion recognition. We will finally discuss promising future research directions of transfer learning for improving the generalizability of automatic emotion recognition systems.

Keywords: transfer learning, generalizability, automatic emotion recognition, speech, image, physiology

OPEN ACCESS

Edited by:
Nicholas Cummins,
University of Augsburg, Germany

Reviewed by:
Ronald Böck,
Otto von Guericke University Magdeburg, Germany
Ziping Zhao,
Tianjin Normal University, China

****Correspondence:***
Kexin Feng
kexin0814@tamu.edu

Specialty section:
This article was submitted to Human-Media Interaction, a section of the journal Frontiers in Computer Science

Received: *31 July 2019*
Accepted: *14 February 2020*
Published: *28 February 2020*

Citation:
Feng K and Chaspari T (2020) A Review of Generalizable Transfer Learning in Automatic Emotion Recognition. Front. Comput. Sci. 2:9. doi: 10.3389/fcomp.2020.00009

1. INTRODUCTION

Emotion plays an important role in human-human or human-computer interaction. Emotionally-aware systems enable the better understanding of human behavior and facilitate uninterrupted and long-term interaction between humans and computers (Beale and Peter, 2008). The recent development of laboratory and real-world sensing systems allows us to fully capture multimodal signal information related to human emotion. This has resulted in a large amount of publicly available datasets with high variability in terms of elicitation methods, speaker demographics, spoken language, and recording conditions. Despite the high availability of such datasets, the amount of data included in each dataset is limited and the emotion-related labels are scarce, therefore prohibiting the reliable training and generalizability of emotion recognition systems. In order to address this challenge, recent studies have proposed transfer learning methods to provide reliable emotion recognition performance, even in unseen contexts, individuals, and conditions (Abdelwahab and Busso, 2018; Lan et al., 2018; Latif et al., 2018; Gideon et al., 2019).

Emerging transfer learning methods can leverage the knowledge from one emotion-related domain to another. The main premise behind such techniques is that people may share similar characteristics when expressing a given emotion. For example, anger may result in increased speech loudness and more intense facial expressions (Siegman and Boyle, 1993). Fear is usually expressed with reduced speech loudness and may produce increased heart rate (Hodges and Spielberger, 1966). These emotion-specific characteristics might be commonly met among people, contributing to the similarity among the various emotional datasets. Therefore, transfer learning approaches can learn common emotion-specific patterns and can be applied across domains for recognizing emotions in datasets with scarce or non-labeled samples. Such techniques can further result in generalizable systems, which can detect emotion for unseen data.

The current manuscript discusses ways in which transfer learning techniques can overcome challenges related to limited amount of data samples, scarce labels, and condition mismatch, and result in robust and generalizable automated systems for emotion recognition. We first introduce

basic concepts in transfer learning (section 2) and discuss the development of major transfer learning methods and their applications in conventional machine learning fields, such as computer vision (section 3). We then review state-of-art work in automatic emotion recognition using transfer learning for speech, image, and physiological modalities (section 4). Finally, we discuss potential promising research directions of transfer learning for improving the generalizability of automatic emotion recognition systems (section 5).

2. BASIC CONCEPTS IN TRANSFER LEARNING

In this section, we will provide the basic definition of transfer learning and discuss several ways to categorize transfer learning methods.

2.1. Definitions

Domain in transfer learning generally refers to a feature space and its marginal probability distribution (Pan et al., 2010). Given a specific domain, a task includes a label space and an objective function that needs to be optimized. Source domain usually refers to a set of data with sufficient data samples, large amount of labels, and potentially high quality (e.g., lab environment). In contrast, the data from a target domain may include limited number of samples and small amount or non-existent labels, and potentially be noisy. Given a source and a target, transfer learning approaches attempt to improve the learning of the target task using knowledge from the source domain.

2.2. Association Metrics Between Source and Target Domains

The selection of the source domain plays an important role in the transfer learning process. A source domain sharing a lot of similarities with the target, is more likely to yield efficient transfer (Pan and Yang, 2009). Similarity can be quantified through the distance between source and target with respect to their data structure (e.g., feature or label distribution), recording conditions (e.g., recording equipment, elicitation methods), and data sample characteristics (e.g., participants with similar demographics, speech samples of same language).

Proposed transfer learning methods typically use a distance metric to maximize the similarity between the source and target domain. Commonly used distance metrics include the: (a) Kullback-Leibler divergence (KL divergence), employing a cross-entropy measure to calculate similarity in the probability distribution between the source and the target domain; (b) Jensen-Shannon divergence (JS divergence), a symmetric version of the KL divergence; (c) Maximum Mean Discrepancy (MMD) and multi-kernel MMD, creating an embedding of the source and target domains on the Reproducing Kernel Hilbert Space (RKHS) and comparing their mean difference; (d) Wasserstein Distance, also known as Earth-Mover (EM) Distance, quantifying the domain difference when there is very little or no overlap between two domain distributions by computing the transport map for every probability density between two domains.

2.3. Categorization of Transfer Learning Techniques

The state-of-art application of transfer learning on automatic emotion tasks can be categorized in two main approaches. The first refers to the availability of labels in the target domain. Supervised transfer learning includes information from labeled data from both the source and target domain during the learning task, while unsupervised learning includes information only from the labels of the source domain (Pan and Yang, 2009). Unsupervised transfer learning enables the design of reliable machine learning systems, even for domains for which labeled data are not available. The second categorization refers to the availability of one or multiple datasets in the source. Single-source transfer learning contains only one dataset, while multi-source transfer learning leverages multiple sets of data in the source domain (Ding et al., 2019).

3. EMERGING WORK ON SUPERVISED AND UNSUPERVISED TRANSFER LEARNING

This section provides an overview of previously proposed methods in the field of transfer learning, summarized into three main categories: (a) statistical-based transfer learning; (b) region selection through domain relevance; and (c) deep transfer learning approaches. We will further discuss these three categories in the following subsections.

3.1. Statistical-Based Transfer Learning

Three types of statistical approaches have been proposed for transfer learning: (a) distribution alignment aims to minimize the shift between the source and target domain to reduce the domain difference; (b) latent space extraction recovers common components between two domains; (c) classifier regularization will increase the ability of regularization for the classifier to predict labels in the target domain.

3.1.1. Alignment Between Source and Target Distributions

The marginal alignment methods aim to find a mapping between source and target distributions. This can be done by setting pairwise constraints between the two domains (Saenko et al., 2010; Kulis et al., 2011). Gopalan et al. (2011) utilized the Grassmann manifold and the incremental learning to search a transformation path between the domains. This method was further improved by Gong et al. (2012), who proposed the Geodesic Flow Kernel (GFK). The Grassmann manifold and Maximum Mean Discrepancy (MMD) are also used in other approaches such as the Domain Invariant Projection (DIP) proposed by Baktashmotlagh et al. (2013). Marginal alignment is one of the first methods appearing in the field of transfer learning, which attempts to find a common distribution between the source and target domains. Despite its promising results, marginal alignment might not always fully align two completely distinct domains to an entirely same distribution, especially

when these include a high mismatch (e.g., different emotional expressions or recording settings).

3.1.2. Latent Space Extraction

The shared subspace extraction methods assume that the feature space of each domain consists of the domain-specific and the domain-invariant components, and comprise one of the most commonly used approaches in transfer learning. These methods attempt to map both the source and target data to a subspace which only keeps the common information between domains to minimize the difference between them. In order to find such a subspace, Pan et al. (2010) proposed the Transfer Component Analysis (TCA) using the Maximum Mean Discrepancy (MMD) and Reproducing Kernel Hilbert Space (RHKS), and significantly reduced the distribution mismatch using a low-dimensional subspace. Additional methods for achieving this goal include boosting [Becker et al., 2013 and Domain-Invariant Component Analysis (DICA); Muandet et al., 2013]. Due to inherent similarities across emotions, the idea of extracting a common latent space has been successfully applied to automatic emotion recognition tasks (Deng et al., 2013; Zheng et al., 2015).

3.1.3. Classifier Regularization

Other approaches have attempted to utilize a regularized version of the classifier trained on the source domain in order to predict the labels in the target domain. Support Vector Machines (SVM) have been widely explored in this process. Yang et al. (2007) proposed the Adapting SVM (A-SVM) which learns a difference function (also referred to as "delta function") between an original and adapted classifier using an objective function similar to the one in the SVM. Other methods include the Projective Model Transfer SVM (PMT-SVM), Deformable Adaptive SVM (DA-SVM), Domain Weighting SVM (DWSVM) (Aytar and Zisserman, 2011), which adapt the weights of a source model to the target domain using various pre-defined constraints (e.g., assign different weights for source and target domain in DWSVM). Bergamo and Torresani (2010) also explored the efficacy of transferring knowledge between different SVM structures using feature Augmentation SVM (AUGSVM), Transductive learning SVM (TSVM) and Domain Transfer SVM (DT-SVM) (Duan et al., 2009). Other types of statistical models, such as maximum entropy classifiers, have been also explored in this process (Daume and Marcu, 2006). Due to its simplicity and efficacy in small data samples, classifier regularization has been applied on various emotion recognition tasks based on physiological signals (Zheng and Lu, 2016).

3.2. Region Selection Through Domain Relevance

The region selection approaches have been mostly introduced in computer vision and rely on the concept of how humans understand a given image. For example, instead of giving equal attention to every part of an image, humans would concentrate more on specific salient objects. Therefore, these approaches aim to identify salient regions of an image by generating a domainness map, and separate the image into a different level of domainness (Tommasi et al., 2016). This domainness feature is further utilized to promote knowledge transfer. Other studies have proposed methods using sparse coding (Long et al., 2013) or abstract auxiliary information (e.g., skeleton or color of an image) (Motiian et al., 2016), which also simulate the way humans comprehend an image as a whole. Hjelm et al. (2018) also utilized the different domain relevance of each part of the image to extract and maximize the mutual information. The region selection methods are very close to the way humans perceive information, and the domainness map makes the methods straightforward and explainable. While similar ideas can also be applied to non-image-related tasks, the determination of domainness levels and validation of the extracted domainness map can be less straightforward.

3.3. Deep Transfer Learning Methods

Deep learning methods are widely explored and applied on transfer learning. Two main types of deep learning approaches have demonstrated promising performance for knowledge transfer: (a) domain adaptation using deep learning, which aims to transfer the knowledge or mitigate the domain difference between source and target with respect to the neural network embedding; and (b) the adversarial and generative learning, which aims to generate data embeddings that are least separable between the source and the target.

3.3.1. Domain Adaptation Using Deep Learning

The large amount of publicly available datasets has yielded several pre-trained deep learning models [e.g., VGG (Simonyan and Zisserman, 2014), VGG-Face (Parkhi et al., 2015), VGG-M-2048 (Chatfield et al., 2014) and AlexNet (Krizhevsky et al., 2012)] which have achieved good performance in image and speech recognition tasks. To address the mismatch between different domains, it is possible to utilize the parameter/structure of pre-trained models to achieve knowledge transfer (e.g., using a model with the same number of hidden layers and same weights learned from the source data). A promising method for achieving this is fine-tuning, which replaces and learns the last layers of the model, while re-adjusting the parameters of the previous ones. A challenge with fine-tuning lies in the fact that the parameters learned on the source task are not preserved after learning the target task. In order to address this "forgetting problem," Rusu et al. (2016) proposed the progressive neural network, which keeps the network trained on the source data, based on which it builds an additional network for the target. Jung et al. (2018) also addressed this problem by keeping the decision boundary unchanged, while also making the feature embeddings extracted for the target close to the ones of the source domain. Utilizing a pre-trained model on the source data includes the following advantages: (a) it speeds up the training process; (b) it potentially increases the ability of generalization, as well as the robustness of the final model; (c) it automatically extracts high-level features between domains.

Although neural network fine-tuning and progressive neural networks can yield benefits to the training process in terms of computational time and ability to generalize (Yosinski et al., 2014), these methods sometimes fail to address the domain difference and may have a poor performance when the source

and target have small overlap. An alternative approach to this has been proposed by Ghifary et al. (2014), who added adaptation layers to the conventional deep learning models and used the Maximum Mean Discrepancy (MMD) for minimizing the domain distribution mismatch between the source and target domains. Instead of making use of a well-trained model, the source data are used in conjunction with the target data in the training process to determine the domain distance. Further research includes determining which layer to be used as the adaptation layer, applying multiple adaptation layers in a model, etc (Tzeng et al., 2014; Long et al., 2015). Moreover, the Joint convolutional neural net (CNN) architecture or Joint MMD (JMMD) (Tzeng et al., 2015; Long et al., 2017) aims to align similar classes between different domains by taking into account the structural information between them.

Various studies on deep transfer learning for automatic emotion recognition have yielded promising results on speech and image datasets (Gideon et al., 2017; Kaya et al., 2017; Abdelwahab and Busso, 2018; Li and Chaspari, 2019. These methods have been less explored for physiological signals, potentially due to the small amount of available data for this modality.

3.3.2. Adversarial and Generative Methods

The idea of adversarial learning for knowledge transfer was proposed by Ganin and Lempitsky (2014) in the domain adversarial neural network (DANN). DANN contains three parts: a feature extractor, a domain classifier, and a task classifier. The feature extractor attempts to learn feature representations which minimize the loss of the task classifier and maximize the loss of the domain classifier. Instead of modifying the loss function based on the distance between two domains, the DANN is able to automatically extract the feature which is common for both domains while maintaining the characteristics of each class (Ganin et al., 2016). Variants of DANNs have further been widely explored. The Domain Separation Network (DSN) proposed by Bousmalis et al. (2016), modified the feature extractor into three encoders (i.e., one for the source, one for the target, and one for both) in order to separate the domain-specific from domain-invariant embeddings. DSN also replaced the domain classifier with a shared decoder, to further ensure that the domain-invariant embedding is useful and can promote the generalizability of the model. According to the multi-adversarial domain adaptation network proposed by Pei et al. (2018), a separate task classifier is trained for every class, which makes different classes less likely to have overlapping distributions. As the number of available datasets increases, networks which handle multiple sources of data are also explored (Xu et al., 2018; Zhao et al., 2018). In order to further avoid the negative transfer, partial transfer learning, which can be done via Bayesian optimization (Ruder and Plank, 2017), is applied in the adversarial neural networks in order to transfer knowledge from large domains to more specific, smaller domains by selecting only part of the source data in the training process (Cao et al., 2018).

Inspired by the two player game, the generative adversarial nets (GAN) were further proposed by Goodfellow et al. (2014) containing a generator and a discriminator. The generator generates fake data from a random distribution and aims to confuse the discriminator, while the discriminator focuses on distinguishing between the real and generated data. In this process, both models can learn from each other and fully explore the patterns of data, since the informed generation of synthetic samples can potentially overcome the mismatch between the source and the target task. Modifications of GAN-based networks are also proposed. For example, Radford et al. (2015) introduced the Deep Convolutional Generative Adversarial Networks (DCGAN) to combine the CNN with GAN. The Wasserstein GAN (WGAN) integrated the Wasserstein distance in the loss function and further improved the training stability (Arjovsky et al., 2017).

The adversarial and generative adversarial neural networks have been successfully applied to speech- and image-based emotion recognition tasks (Wang and Zheng, 2015; Motiian et al., 2017; Sun et al., 2018) with promising results.

4. TRANSFER LEARNING FOR AUTOMATIC EMOTION RECOGNITION

In this section, we discuss the application of transfer learning on three modalities commonly used in the automatic emotion recognition task: (a) speech; (b) video (or image); and (c) physiology. Sentiment analysis is not included in this manuscript, since it is related to crowd-sourced data, which are beyond the scope of the review.

4.1. Transfer Learning for Speech-Based Emotion Recognition

Because of the multi-faceted information included in the speech signal, transfer learning has been widely applied in speech-based emotion recognition (**Table 1**). Previously proposed approaches attempt to transfer the knowledge between datasets collected under similar conditions (e.g., audio signals collected by actors in the lab) (Abdelwahab and Busso, 2015, 2018; Sagha et al., 2016; Zhang et al., 2016; Deng et al., 2017; Gideon et al., 2017; Neumann and Vu, 2019) or using the knowledge from acted in-lab audio signals to spontaneous speech collected in-the-wild (Deng et al., 2014b; Mao et al., 2016; Zong et al., 2016; Song, 2017; Gideon et al., 2019; Li and Chaspari, 2019).

Different types of transfer learning architectures have been explored in speech-based emotion recognition, including the statistical methods (Deng et al., 2013, 2014a,c; Abdelwahab and Busso, 2015; Song et al., 2015; Sagha et al., 2016; Zong et al., 2016; Song, 2017), the adversarial or generative networks (Chang and Scherer, 2017; Abdelwahab and Busso, 2018; Gideon et al., 2019; Latif et al., 2019), and other neural network structures (Mao et al., 2016; Deng et al., 2017; Gideon et al., 2017; Li and Chaspari, 2019; Neumann and Vu, 2019; Zhou and Chen, 2019). A commonly used input of the aforementioned approaches includes the feature set proposed by the INTERSPEECH emotion challenge and INTERSPEECH paralinguistic challenges (Schuller et al., 2009b, 2010, 2013), which typically contains the first 12 Mel Frequency Cepstral Coefficients, root-mean-square energy, zero-crossing rate, voice probability, and fundamental frequency (Deng et al.,

TABLE 1 | Overview of previously proposed transfer learning methods for speech-based emotion recognition.

References	Dataset	In-lab/real-world transfer learning	Acted/ spontaneous transfer learning	Emotional labels	Cross-linguistic transfer learning	Type of transfer learning	Input features
Deng et al., 2013	**source:** TUM AVIC (Schuller et al., 2009a), EMO-DB (Burkhardt et al., 2005), eNTERFACE (Martin et al., 2006), SUSAS (Hansen and Bou-Ghazale, 1997), VAM (Grimm et al., 2008) **target:** FAU AEC (Steidl, 2009)	In-lab; real-world	Acted & spontaneous	Valence	English German	Autoencoder for aligning source to target SVM for classification	INTERSPEECH 2009 emotion challenge
Deng et al., 2014b	**source:** SUSAS (Hansen and Bou-Ghazale, 1997), ABC (Schuller et al., 2007) **target:** FAU AEC (Steidl, 2009)	**source:** In-lab; real-world **target:** Real-world	**source:** Acted; spontaneous **target:** Spontaneous	Valence	English German	Adaptive denoising autoencoder (DAE)	INTERSPEECH 2009 emotion challenge
Deng et al., 2014c	**source:** SUSAS (Hansen and Bou-Ghazale, 1997), ABC (Schuller et al., 2007) **target:** FAU AEC (Steidl, 2009)	**source:** In-lab; real-world **target:** Real-world	**source:** Acted; spontaneous **target:** Spontaneous	Valence	English German	Encoders trained separately for domains one layer nn maps subspace to target	INTERSPEECH 2009 emotion challenge
Deng et al., 2014a	**source:** SUSAS (Hansen and Bou-Ghazale, 1997), ABC (Schuller et al., 2007) **target:** FAU AEC (Steidl, 2009)	**source:** In-lab; real-world **target:** Real-world	**source:** Acted; spontaneous **target:** Spontaneous	Valence	English German	Shared-hidden-layer autoencoder. A common encoder which also aims to minimize reconstruction error	INTERSPEECH 2009 emotion challenge
Song et al., 2015	**source/target:** EMO-DB (Burkhardt et al., 2005), eNTERFACE (Martin et al., 2006)	In-lab	Acted	Angry, disgusted, fear, happy, sad	English German	Transfer principal component analysis and sparse coding based method	INTERSPEECH 2010 paralinguistic challenge
Abdelwahab and Busso, 2015	**source:** IEMOCAP (Busso et al., 2008), SEMAINE (McKeown et al., 2011) **target:** RECOLA (Ringeval et al., 2013)	In-lab	Acted & spontaneous	Arousal and valence	English French	Domain adaptation for SVM incremental adaptation for SVM	INTERSPEECH 2011 speaker state feature
Mao et al., 2016	**source:** EMO-DB (Burkhardt et al., 2005), ABC (Schuller et al., 2007) **target:** FAU AEC (Steidl, 2009)	**source:** In-lab **target:** Real-world	**source:** Acted **target:** Spontaneous	Valence	English German	Sharing priors between related source and target classes	INTERSPEECH 2009 emotion challenge
Sagha et al., 2016	**source/target:** EMO-DB (Burkhardt et al., 2005), SAVEE (Haq et al., 2008), EMOVO (Costantini et al., 2014), Polish (Staroniewicz and Majewski, 2009)	In-lab	Acted	Valence	English, German, Italian, Polish	Kernel canonical correlation analysis (KCCA)	INTERSPEECH 2009 emotion challenge
Zhang et al., 2016	**source:** RAVDESS (Livingstone and Russo, 2018) **target:** UMSSED (Zhang et al., 2015)	In-lab	Acted	Angry, happy neutral, sad	No	Multi-task learning	INTERSPEECH computational paralinguistics challenge 2013
Zong et al., 2016	**source/target:** EMO-DB (Burkhardt et al., 2005), eNTERFACE (Martin et al., 2006), AFEW (Dhall et al., 2012)	In-lab & real-world	Acted & spontaneous	Angry, disgusted, afraid happy, neutral, sad	English German	Domain-adaptive least-squares regression (DaLSR)	INTERSPEECH 2009 emotion challenge
Song, 2017	**source/target:** EMO-DB (Burkhardt et al., 2005), eNTERFACE (Martin et al., 2006), FAU AEC (Steidl, 2009)	In-lab & real-world	Acted & spontaneous	Angry, disgusted, afraid, happy, sad	English German	Linear subspace learning	INTERSPEECH 2010 paralinguistic challenge
Deng et al., 2017	**source/target:** EMO-DB (Burkhardt et al., 2005), GeWEC (Bänziger and Scherer, 2010)	In-lab	Acted	Categorical emotions	German French	Denoising autoencoder	INTERSPEECH 2009 emotion challenge

(Continued)

TABLE 1 | Continued

References	Dataset	In-lab/real-world transfer learning	Acted/ spontaneous transfer learning	Emotional labels	Cross-linguistic transfer learning	Type of transfer learning	Input features
				Valence and arousal	Shared-hidden-layer autoencoder extreme learning machine autoencoder		
Gideon et al., 2017	**source/target:** IEMOCAP (Busso et al., 2008), MSP-IMPROV (Busso et al., 2016)	In-lab	Acted	Angry, neutral, sad, happy	No	Progressive neural network (PNN)	Geneva minimalistic acoustic parameter set (GeMAPS)
Chang and Scherer, 2017	**source:** AMI (Carletta et al., 2005) **target:** IEMOCAP (Busso et al., 2008)	**source:** Real-world **target:** In-lab	Acted & spontaneous	Valence and activation	No	Deep convolutional generative adversarial networks (DCGAN)	Speech spectrogram
Abdelwahab and Busso, 2018	**source:** IEMOCAP (Busso et al., 2008), MSP-IMPROV (Busso et al., 2016) **target:** MSP-Podcast (Lotfian and Busso, 2017)	In-lab	Acted & spontaneous	Arousal, valence, dominance	No	Domain adversarial neural network (DANN)	INTERSPEECH computational paralinguistics challenge 2013
Gideon et al., 2019	**source/target:** IEMOCAP (Busso et al., 2008), MSP-IMPROV (Busso et al., 2016) PRIORI Emotion (Khorram et al., 2018)	In-lab & real-world	Acted & spontaneous	Valence	No	Adversarial discriminative domain generalization (ADDoG)	Mel Filter Bank (MFB)
Li and Chaspari, 2019	**source:** IEMOCAP (Busso et al., 2008), CREMA-D (Cao et al., 2014), RAVDESS (Livingstone and Russo, 2018), eNTERFACE (Martin et al., 2006) **target:** IEMOCAP (Busso et al., 2008)	In-lab	**source:** Acted **target:** Spontaneous	Angry, happy, sad, afraid	No	Feedforward neural network fine-tuning progressive neural network (PNN)	INTERSPEECH 2009 emotion challenge
Neumann and Vu, 2019	**source/target:** IEMOCAP (Busso et al., 2008), MSP-IMPROV (Busso et al., 2016)	In-lab	Acted	Angry, happy, sad, neutral	No	A latent feature space was learned on source domain using an encoder-decoder. Such space was added as feature vector in attentive convolutional neural network	MFCC feature
Latif et al., 2019	**source/target:** EMO-DB (Burkhardt et al., 2005), SAVEE (Jackson and Haq, 2014), EMOVO (Costantini et al., 2014), URDC (Latif et al., 2018)	In-lab & real-world	Acted & spontaneous	Positive/ negative valence	German, Urdu Italian, English	Similar to GAN structure but source data was used instead of generated fake data.	Geneva minimalistic acoustic parameter set (GeMAPS)
Zhao et al., 2019	**source:** eGender (Burkhardt et al., 2010) **target:** EMO-DB (Burkhardt et al., 2005), IEMOCAP (Busso et al., 2008)	**source:** Real-world **target:** In-lab	**source:** spontaneous **target:** acted	Continuous prediction or classification: neutral, happiness, sadness, anger	English German	Learn age and gender attributes separately then transfer these knowledge by feeding such information to emotion model.	INTERSPEECH 2010 configuration

(Continued)

TABLE 1 | Continued

References	Dataset	In-lab/real-world transfer learning	Acted/ spontaneous transfer learning	Emotional labels	Cross-linguistic transfer learning	Type of transfer learning	Input features
Zhou and Chen, 2019	**source:** Aibo-Ohm and Aibo-Mont (Steidl, 2009) **target:** EMO-DB (Burkhardt et al., 2005)	In-lab	**source:** Spontaneous **target:** acted	Binary negative / positive	No	Data was relabeled to reveal domain info class-wise adversarial domain adaptation two stages training: train encoder, predictor fix predictor, train encoder only	Geneva minimalistic acoustic parameter set (GeMAPS)

2013, 2014b, 2017; Mao et al., 2016; Sagha et al., 2016; Zhang et al., 2016; Zong et al., 2016; Song, 2017; Abdelwahab and Busso, 2018; Li and Chaspari, 2019; Zhao et al., 2019).

Statistical values of these descriptors, including maximum, minimum, range, time position of the maximum and minimum, average, standard deviation, skewness, kurtosis, as well as the first- and second-order coefficient of a linear regression model are extracted from the frame-based measures. Other approaches also include the speech spectrogram as an input to convolutional-based neural networks (Gideon et al., 2019). Previously proposed transfer learning methods for speech emotion recognition employ same classes for the source and target data. Two commonly used baseline methods against which the proposed transfer learning approaches are compared include in-domain training and out-of-domain training (only used data from source domain). The first performs training and testing by solely using labeled data from target domain, while the second trains the model on the source data and tests on the target. Results indicate that the proposed transfer learning methods outperform the out-of-domain methods, and are equivalent to or sometimes surpass in-domain training, indicating the potential of leveraging multiple sources of emotion-specific speech data to improve emotion recognition performance.

Besides speech data, audio signals from music clips have been also applied for emotion recognition (Zhang et al., 2016). However, because of the limited number of emotion-based datasets with music signals, as well as the significant domain mismatch between music and speech, this application is relatively less explored.

4.2. Transfer Learning for Video/Image-Based Emotion Recognition

Facial expressions convey a rich amount of information related to human emotion. A variety of transfer learning techniques have been explored for video/image-based automatic emotion recognition (**Table 2**). The state-of-art transfer learning approach to the video-based emotion recognition includes obtaining high-level features using mainly a convolutional neural network (CNN) trained on large sources of data (e.g., VGG; Simonyan and Zisserman, 2014) (Kaya et al., 2017; Aly and Abbott, 2019; Ngo and Yoon, 2019, or transfering the knowledge from higher-quality auxiliary image datasets (e.g., skeleton or color of an image, image with description text) (Xu et al., 2016). Source datasets in this case might not necessarily contain the same labeled classes as the target dataset. Occluded facial images, which are common in daily life, are also utilized to help with the generalization and robustness of the overall system (Xu et al., 2015). More advanced transfer learning approaches, such as adversarial methods, are less explored in this process. A possible reason is that the high-level image features are relatively easier to obtain and knowledge from other domains might not be able to significant help. Another reason could be the selection of source domain is more important for facial emotion recognition (Sugianto and Tjondronegoro, 2019). In order to recognize emotions from video clips, every frame of the clip is analyzed and the final decision is made based on voting methods, such as major voting on the separate frames (Zhang et al., 2016). Face detection methods, such as the deformable parts model (DPM) (Mathias et al., 2014), may also be used to avoid the influence of irrelevant regions of the video frame (Kaya et al., 2017).

4.3. Transfer Learning for Emotion Recognition Based on Physiological Signals

A small amount of previous work has attempted to perform transfer learning on physiological signals for emotion recognition (**Table 3**). Among the various physiological signals, the electroencephalogram (EEG) is the most commonly used in transfer learning, probably due to the rich amount of information included in this signal. Because of the limited number of datasets including physiological signals, as well as the high variability across people, the knowledge transfer between different datasets is less efficient and relatively less explored. Commonly used transfer learning applications attempt to train subject-specific (personalized) models by providing knowledge learned from subjects similar to the one in the test (Lin and Jung, 2017; Lin, 2019), or simply consider all the members in the group by assigning different weights (Li et al., 2019). Other methods include statistical approaches, such as Principal

TABLE 2 | Overview of the previous work on transfer learning for video/image-based emotion recognition.

References	Dataset	In-lab/real-world transfer learning	Acted/ spontaneous transfer learning	Same labels between source & target	Emotional labels	Type of transfer learning
Ng et al., 2015	**source:** VGG (Simonyan and Zisserman, 2014) AlexNet (Krizhevsky et al., 2012), FER-2013 (Goodfellow et al., 2013) **target:** EmotiW 2015 (Dhall et al., 2015)	Real-world	Spontaneous	No for VGG/AlexNet yes for FER-2013	Neutral, angry, disgusted, sad, fear, happy, surprised	Two-stage fine-tuning based on VGG/AlexNet and the target data
Xu et al., 2015	**source:** MSRA-CFW (Zhang et al., 2012) **target:** self-built database contains CK+ (Lucey et al., 2010), JAFFE (Lyons et al., 1999), KDEF (Goeleven et al., 2008), PICS (PIC, 2013)	Real-world	Spontaneous	No	Neutral, angry, disgusted, sad, fear, happy, surprised	Feature transfer by training two facial Identification convolutional networks
Xu et al., 2016	**source:** Flickr (Borth et al., 2013) **target:** YouTube (Jiang et al., 2014) Ekman-6 emotion dataset	Real-world	Spontaneous	Yes	8 Primary emotions 24 primary & Secondary emotions	Auxiliary image transfer encoding Using auxiliary data (e.g., image with description text)
Kaya et al., 2017	**source:** VGG-Face (Parkhi et al., 2015), VGG-M-2048 (Chatfield et al., 2014) FER-2013 (Goodfellow et al., 2013) **target:** EmotiW 2015 (Dhall et al., 2015), EmotiW 2016 (Dhall et al., 2016), CK+ (Lucey et al., 2010), MMI (Valstar and Pantic, 2010), RECOLA (Ringeval et al., 2013), First impressions challenge (Escalante et al., 2016)	Real-world	Spontaneous	Yes	Neutral, angry, disgusted, sad, fear, happy, surprised	Fine-tuning during convolutional Neural network (CNN) training
Ngo and Yoon, 2019	**source:** ResNet-50(He et al., 2016) **target:** AffectNet (Mollahosseini et al., 2017)	Real-world	Spontaneous	No	Neutral, happiness, sadness, surprise, fear, disgust, anger, contempt	Fine-tuning on the well-trained ResNet-50 net
Aly and Abbott, 2019	**source:** AlexNet (Krizhevsky et al., 2012), JAFEE (Lyons et al., 1998) CK+ (Lucey et al., 2010) **target:** VT-KFER (Aly et al., 2015), 300W (Sagonas et al., 2016)	VTKFER: in-lab 300W: real-world	VTKFER: acted 300W: spontaneous	Yes	Happiness, sadness, surprise, disgust, fear, anger	Multi-stage Progressive Transfer Learning (MSPTL) fine tune the AlexNet in multiple stages using different data (simple to more challenging or non-frontal)
Sugianto and Tjondronegoro, 2019	**source:** ResNet-50(He et al., 2016) MS-CELEB-1M(He et al., 2016) VGGFace2(He et al., 2016) CK+ (Lucey et al., 2010) **target:** AffectNet (Mollahosseini et al., 2017)	Real-world	Spontaneous	No	Neutral, happiness, sadness, surprise, fear, disgust, anger, contempt	Fine-tuning on CK+ (relevant domain) lowers the performance due to the large knowledge gap. General to specific knowledge transfer performs best.

Component Analysis (PCA) and adaptive subspace feature matching (ASFM) (Chai et al., 2017). As the development of wearable devices progresses and more physiological data related to emotional experiences become available, transfer learning methods appear to depict a great potential in this domain.

TABLE 3 | Overview of transfer learning methods for emotion recognition based on physiological signals.

References	Dataset	Elicitation method	In-lab Real-world	Emotional labels	Type of transfer learning
Zheng et al., 2015	SEED (Zheng and Lu, 2015)	Video	In-lab	Three emotions (positive, neutral, and negative)	Personalized transfer learning Transfer component analysis (TCA) kernel principal component analysis (KPCA)
Chai et al., 2016	SEED (Zheng and Lu, 2015)	Video	In-lab	Three emotions (positive, neutral, and negative)	Subspace alignment auto-encoder
Zheng and Lu, 2016	SEED (Zheng and Lu, 2015)	Video	In-lab	Three emotions (positive, neutral, and negative)	Transductive parameter transfer (TPT) Transductive SVM (T-SVM) Transfer component analysis (TCA) Kernel PCA (KPCA)
Lin and Jung, 2017	Oscar soundtrack EEG dataset (Lin et al., 2010)	Music	In-lab	Valence and arousal	Conditional transfer learning framework to determine How transferable is a model to a given individual
Chai et al., 2017	SEED (Zheng and Lu, 2015)	Video	In-lab	Three emotions (positive, neutral, and negative)	Adaptive subspace feature matching
Lan et al., 2018	SEED (Zheng and Lu, 2015) DEAP (Koelstra et al., 2011)	Video	In-lab	Three emotions (positive, neutral, and negative)	Transfer component analysis (TCA) Geodesic flow kernel (GFK) Domain adaptation Kernel principal component Analysis (KPCA)
Lin, 2019	MDME (Lin et al., 2015) SDMN (Lin et al., 2010)	Music	In-lab	Binary valence and arousal	Principal component analysis (RPCA)-embedded transfer learning personalized cross-day model. Use Riemannian distance and RPCA to select similar samples within dataset
Li et al., 2019	SEED (Zheng and Lu, 2015)	Video	In-lab	Three emotions (positive, neutral, and negative)	Known objects have seperate classifiers such classifiers were ensembled using style transfer mapping (STM) method for a new object.
Zhang et al., 2019	SEED (Zheng and Lu, 2015)	Video	In-lab	Three emotions (positive, neutral, and negative)	A CNN is used as feature extractor from Electrodes-frequency Distribution Maps. Deep domain confusion (DDC) narrowed feature difference between domains. EFDMs and CNN are used for classification.

TABLE 4 | Overview of current research.

Signal	Common transfer learning method	Overview
Speech	Statistical-based transfer learning Deep transfer learning methods	Speech signal has been widely used for emotion recognition, both statistical and deep learning methods are widely explored.
Video/image	Deep transfer learning methods	Motivated by the wide adoption of deep learning in image processing, such methods are widely used for emotion recognition in the last years.
Physiological	Statistical-based transfer learning	Neural networks were not widely used for physiological signals in recent years, but researchers have started to apply deep transfer learning methods.

5. DISCUSSION

In this section, we will provide a brief summary of the current application of transfer learning on automatic emotion recognition and outline potential aspects for future research.

5.1. Summary of Current Research

Previous work has explored transfer learning for automatic emotion recognition in the three commonly used signals (i.e., speech, video, and physiology) (**Table 4**). Transfer learning for speech aims to transfer the knowledge between the different datasets using the state-of-art transfer learning methods, such as adversarial or generative networks. Image-based transfer learning is mainly utilized to extract high-level features from images or their auxiliary data (e.g., image description text) using convolutional neural networks (CNN). For physiological signals, transfer learning has promoted the design of personalized models through statistical methods, though the data scarcity and high inter-individual variability yield highly variable results across subjects. Speech and video signals allow for the design of sophisticated systems that can detect multiple emotions (i.e., up to seven emotions), while physiological data usually yield more coarse-grained emotion recognition.

5.2. Potential Future Directions

5.2.1. Multi-Modal Transfer Learning for Emotion Recognition

Multi-modal sources of data have been widely used in automatic emotion recognition tasks in order to supplement information

from multiple modalities and reduce potential bias from one single signal (Busso et al., 2004; Wöllmer et al., 2010). However, transfer learning methods play a very limited role in this process. Common knowledge transfer in multi-modal methods include fine-tune well trained models to a specific type of signal (Vielzeuf et al., 2017; Yan et al., 2018; Huang et al., 2019; Ortega et al., 2019), or fine-tune different well-trained models to both speech and video signals (Ouyang et al., 2017; Zhang et al., 2017; Ma et al., 2019). Other usage of transfer learning for multi-modal methods includes leveraging the knowledge from one signal to another (e.g., video to speech) to reduce the potential bias (Athanasiadis et al., 2019). While these methods provide a promising performance, applying the more recent transfer learning methods, or utilizing more types of signals and leveraging knowledge between multiple signals, is likely to improve transfer learning and boost emotion recognition accuracy.

5.2.2. Transferring Emotionally-Relevant Knowledge Between In-lab and Real-Life Conditions

Most of the current studies focus on transferring knowledge between datasets collected in the lab, due to the relative less availability of real-world dataset, especially for the speech or physiological signals. (Abdelwahab and Busso, 2015, 2018; Zheng et al., 2015; Chai et al., 2016, 2017; Sagha et al., 2016; Zhang et al., 2016; Zheng and Lu, 2016; Gideon et al., 2017; Lin and Jung, 2017; Lan et al., 2018; Li and Chaspari, 2019). While some of these datasets try to simulate a naturalistic scenarios, they are still very different from actual real-world conditions, since they contain less noisy data, collected in high quality conditions (e.g., no occluded video or far-field speech), and might not be able to successfully elicit all possible emotions met in real-life conditions (e.g., grief). Recently, there have been multiple efforts to collect emotion datasets in the wild, such as the PRIORI (Khorram et al., 2018) and the AVEC In-The-Wild Emotion Recognition Challenge (Dhall et al., 2015). Exploring the ability of transferring knowledge from data collected in the lab to data obtained in real-life conditions can significantly extend the applicability of emotion recognition applications in real-life (e.g., quantifying well-being, tracking mental health indices; Khorram et al., 2018).

5.2.3. Multi-Source Transfer Learning

With the advent of a large number of emotion recognition corpora, multi-source transfer learning methods provide a promising research direction. By leveraging the variability from multiple data sources collected under different contextual and recording conditions, multi-source transfer learning might be able to provide highly robust and generalizable systems. Multi-source transfer learning can also lay a foundation for modeling various aspects of human behavior different than emotion (e.g., mood, anxiety), where only a limited number of datasets with a small number of data samples are available. Multi-source transfer learning has not been explored for automatic emotion recognition task, which also makes it a great future research direction.

5.2.4. Distinct Labels Between the Source and Target Domains

A potential challenge in automatic emotion recognition lies in the fact that the various datasets might include different types of emotional labels, which can introduce high mismatch between the source and target domains. Especially in the case where the source domain includes data collected in the lab, the corresponding labels mostly only include basic emotions (e.g., happy, anger, fear, surprise, and neutral). However, the emotional classes in the target domain might be slightly different, since they might include subtle real-world emotions, such as frustration, disapproval, etc. Understanding and modeling associations between primary and secondary emotions can potentially contribute toward more accurate emotion inferences in real-life.

5.2.5. Transfer Learning for Cross-Cultural and Cross-Linguistic Emotion Recognition

Emotions can be expressed in different ways across different cultures and languages. For example, emotions may be expressed in a direct and noticeable way in Western countries, while emotional expression tends to be more subtle in parts of Asia (Davis et al., 2012; Gökçen et al., 2014). Even though previous studies have explored the knowledge transfer across European languages (e.g., German, French, Italian, and Polish) (Sagha et al., 2016), indicating that languages are not a key factor for automatic emotion recognition, extensive experiments with non-Western languages/cultures might be able to provide additional insights for advancing the field of transfer learning in emotion recognition. At the same time, most emotional datasets include Caucasian subjects, while a few samples of collected data contain participants from different ethnicities and races (Schuller et al., 2009a). It would be beneficial to examine potential discrepancies related to the linguistic speaking style and facial expressions for building generalizable emotion recognition systems across cultures.

6. CONCLUSION

In this manuscript, we reviewed emerging methods in supervised and unsupervised transfer learning, as well as successful applications and promising future research directions of transfer learning for automatic emotion recognition. We first provided an overview of basic transfer learning methods mostly used in image and speech processing, including statistical approaches, deep transfer learning, and region selection through domain relevance. We then expanded upon transfer learning applications for emotion recognition studying three main modalities of speech, image, and physiological signals. Findings from previous work suggest the feasibility of transfer learning approaches in building reliable emotion recognition systems, yielding improved performance compared to in-domain learning (i.e., training and testing models on samples from the same dataset). Despite the encouraging findings, various implications for future work exist by leveraging multiple sources and modalities of emotional data, which have the potential

to yield transferrable emotional embeddings toward novel computational models of human emotion, and human behavior in general.

AUTHOR CONTRIBUTIONS

KF made contributions to the conception, design, and analysis of existing work, and drafted the research article. TC contributed to the conception and design of the work, and revised the article.

FUNDING

This research was funded by the Engineering Information Foundation (EiF18.02) and the Texas A&M Program to Enhance Scholarly and Creative Activities (PESCA).

REFERENCES

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

Abdelwahab, M., and Busso, C. (2018). Domain adversarial for acoustic emotion recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* 26, 2423–2435. doi: 10.1109/TASLP.2018.2867099

Aly, S., Trubanova, A., Abbott, L., White, S., and Youssef, A. (2015). "Vt-kfer: a kinect-based rgbd+ time dataset for spontaneous and non-spontaneous facial expression recognition," in *2015 International Conference on Biometrics (ICB)* (Phuket: IEEE), 90–97.

Aly, S. F., and Abbott, A. L. (2019). "Facial emotion recognition with varying poses and/or partial occlusion using multi-stage progressive transfer learning," in *Scandinavian Conference on Image Analysis* (Norrköping: Springer), 101–112.

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein GAN. *arXiv: 1701.07875*.

Athanasiadis, C., Hortal, E., and Asteriadis, S. (2019). Audio-visual domain adaptation using conditional semi-supervised generative adversarial networks. *Neurocomputing*. doi: 10.1016/j.neucom.2019.09.106

Aytar, Y., and Zisserman, A. (2011). "Tabula rasa: model transfer for object category detection," in *2011 International Conference on Computer Vision* (Barcelona: IEEE), 2252–2259. doi: 10.1109/ICCV.2011.6126504

Baktashmotlagh, M., Harandi, M. T., Lovell, B. C., and Salzmann, M. (2013). "Unsupervised domain adaptation by domain invariant projection," in *Proceedings of the IEEE International Conference on Computer Vision* (Sydney, NSW), 769–776. doi: 10.1109/ICCV.2013.100

Bänziger, T., and Scherer, K. R. (2010). "Introducing the geneva multimodal emotion portrayal (GEMEP) corpus," in *Blueprint for Affective Computing: A Sourcebook*, eds K. R. Scherer, T. Bänziger, and E. B. Roesch (Oxford: Oxford University Press), 271–294.

Beale, R., and Peter, C. (Eds.). (2008). "The role of affect and emotion in HCI," in *Affect and Emotion in Human-Computer Interaction* (Berlin; Heidelberg: Springer), 1–11.

Becker, C. J., Christoudias, C. M., and Fua, P. (2013). "Non-linear domain adaptation with boosting," in *Advances in Neural Information Processing Systems* (Lake Tahoe, BC), 485–493.

Bergamo, A., and Torresani, L. (2010). "Exploiting weakly-labeled web images to improve object classification: a domain adaptation approach," in *Advances in Neural Information Processing Systems* (Vancouver, CA), 181–189.

Borth, D., Ji, R., Chen, T., Breuel, T., and Chang, S.-F. (2013). "Large-scale visual sentiment ontology and detectors using adjective noun pairs," in *Proceedings of the 21st ACM International Conference on Multimedia* (Barcelona: ACM), 223–232. doi: 10.1145/2502081.2502282

Bousmalis, K., Trigeorgis, G., Silberman, N., Krishnan, D., and Erhan, D. (2016). "Domain separation networks," in *Advances in Neural Information Processing Systems* (Barcelona), 343–351.

Burkhardt, F., Eckert, M., Johannsen, W., and Stegmann, J. (2010). "A database of age and gender annotated telephone speech," in *LREC* (Malta).

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

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., et al. (2008). IEMOCAP: interactive emotional dyadic motion capture database. *Lang. Resour. Evaluat.* 42:335. doi: 10.1007/s10579-008-9076-6

Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., et al. (2004). "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proceedings of the 6th International Conference on Multimodal Interfaces* (State College, PA: ACM), 205–211. doi: 10.1145/1027933.1027968

Busso, C., Parthasarathy, S., Burmania, A., AbdelWahab, M., Sadoughi, N., and Provost, E. M. (2016). MSP-IMPROV: an acted corpus of dyadic interactions to study emotion perception. *IEEE Trans. Affect. Comput.* 8, 67–80. doi: 10.1109/TAFFC.2016.2515617

Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., and Verma, R. (2014). CREMA-D: crowd-sourced emotional multimodal actors dataset. *IEEE Trans. Affect. Comput.* 5, 377–390. doi: 10.1109/TAFFC.2014.2336244

Cao, Z., Long, M., Wang, J., and Jordan, M. I. (2018). "Partial transfer learning with selective adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT), 2724–2732. doi: 10.1109/CVPR.2018.00288

Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., et al. (2005). "The AMI meeting corpus: a pre-announcement," in *International Workshop on Machine Learning for Multimodal Interaction* (Edinburgh: Springer), 28–39.

Chai, X., Wang, Q., Zhao, Y., Li, Y., Liu, D., Liu, X., et al. (2017). A fast, efficient domain adaptation technique for cross-domain electroencephalography (EEG)-based emotion recognition. *Sensors* 17:1014. doi: 10.3390/s17051014

Chai, X., Wang, Q., Zhao, Y., Liu, X., Bai, O., and Li, Y. (2016). Unsupervised domain adaptation techniques based on auto-encoder for non-stationary eeg-based emotion recognition. *Comput. Biol. Med.* 79, 205–214. doi: 10.1016/j.compbiomed.2016.10.019

Chang, J., and Scherer, S. (2017). "Learning representations of emotional speech with deep convolutional generative adversarial networks," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (New Orleans, LA: IEEE), 2746–2750. doi: 10.1109/ICASSP.2017.7952656

Chatfield, K., Simonyan, K., Vedaldi, A., and Zisserman, A. (2014). Return of the devil in the details: delving deep into convolutional nets. *arXiv: 1405.3531*. doi: 10.5244/C.28.6

Costantini, G., Iaderola, I., Paoloni, A., and Todisco, M. (2014). "Emovo corpus: an Italian emotional speech database," in *International Conference on Language Resources and Evaluation (LREC 2014)* (Reykjavik: European Language Resources Association), 3501–3504.

Daume, H. III., and Marcu, D. (2006). Domain adaptation for statistical classifiers. *J. Artif. Intell. Res.* 26, 101–126. doi: 10.1613/jair.1872

Davis, E., Greenberger, E., Charles, S., Chen, C., Zhao, L., and Dong, Q. (2012). Emotion experience and regulation in china and the united states: how do culture and gender shape emotion responding? *Int. J. Psychol.* 47, 230–239. doi: 10.1080/00207594.2011.626043

Deng, J., Frühholz, S., Zhang, Z., and Schuller, B. (2017). Recognizing emotions from whispered speech based on acoustic feature transfer learning. *IEEE Access* 5, 5235–5246. doi: 10.1109/ACCESS.2017.2672722

Deng, J., Xia, R., Zhang, Z., Liu, Y., and Schuller, B. (2014a). "Introducing shared-hidden-layer autoencoders for transfer learning and their application in acoustic emotion recognition," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Florence: IEEE), 4818–4822.

Deng, J., Zhang, Z., Eyben, F., and Schuller, B. (2014b). Autoencoder-based unsupervised domain adaptation for speech emotion recognition. *IEEE Signal Process. Lett.* 21, 1068–1072. doi: 10.1109/LSP.2014.2324759

Deng, J., Zhang, Z., Marchi, E., and Schuller, B. (2013). "Sparse autoencoder-based feature transfer learning for speech emotion recognition," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction* (Geneva: IEEE), 511–516.

Deng, J., Zhang, Z., and Schuller, B. (2014c). "Linked source and target domain subspace feature transfer learning–exemplified by speech emotion recognition," in *2014 22nd International Conference on Pattern Recognition* (Stockholm: IEEE), 761–766.

Dhall, A., Goecke, R., Joshi, J., Hoey, J., and Gedeon, T. (2016). "Emotiw 2016: video and group-level emotion recognition challenges," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction* (Tokyo: ACM), 427–432.

Dhall, A., Goecke, R., Lucey, S., and Gedeon, T. (2012). Collecting large, richly annotated facial-expression databases from movies. *IEEE Multimedia* 19, 34–41. doi: 10.1109/MMUL.2012.26

Dhall, A., Ramana Murthy, O., Goecke, R., Joshi, J., and Gedeon, T. (2015). "Video and image based emotion recognition challenges in the wild: emotiw 2015," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction* (Seattle, WA: ACM), 423–426.

Ding, Z., Zhao, H., and Fu, Y. (2019). *Multi-source Transfer Learning*. Cham: Springer International Publishing.

Duan, L., Tsang, I. W., Xu, D., and Maybank, S. J. (2009). "Domain transfer SVM for video concept detection," in *2009 IEEE Conference on Computer Vision and Pattern Recognition* (Miami, FL: IEEE), 1375–1381.

Escalante, H. J., Ponce-López, V., Wan, J., Riegler, M. A., Chen, B., Clapés, A., et al. (2016). "Chalearn joint contest on multimedia challenges beyond visual analysis: an overview," in *2016 23rd International Conference on Pattern Recognition (ICPR)* (Cancún: IEEE), 67–73.

Ganin, Y., and Lempitsky, V. (2014). Unsupervised domain adaptation by backpropagation. *arXiv: 1409.7495*.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., et al. (2016). Domain-adversarial training of neural networks. *J. Mach. Learn. Res.* 17, 2096–2030.

Ghifary, M., Kleijn, W. B., and Zhang, M. (2014). "Domain adaptive neural networks for object recognition," in *Pacific Rim International Conference on Artificial Intelligence* (Queensland: Springer), 898–904.

Gideon, J., Khorram, S., Aldeneh, Z., Dimitriadis, D., and Provost, E. M. (2017). Progressive neural networks for transfer learning in emotion recognition. *arXiv: 1706.03256*. doi: 10.21437/Interspeech.2017-1637

Gideon, J., McInnis, M. G., and Provost, E. M. (2019). Barking up the right tree: improving cross-corpus speech emotion recognition with adversarial discriminative domain generalization (ADDoG). *arXiv: 1903.12094*. doi: 10.1109/TAFFC.2019.2916092

Goeleven, E., De Raedt, R., Leyman, L., and Verschuere, B. (2008). The Karolinska directed emotional faces: a validation study. *Cogn. Emot.* 22, 1094–1118. doi: 10.1080/02699930701626582

Gökçen, E., Furnham, A., Mavroveli, S., and Petrides, K. (2014). A cross-cultural investigation of trait emotional intelligence in Hong Kong and the UK. *Pers. Individ. Diff.* 65, 30–35. doi: 10.1016/j.paid.2014.01.053

Gong, B., Shi, Y., Sha, F., and Grauman, K. (2012). "Geodesic flow kernel for unsupervised domain adaptation," in *2012 IEEE Conference on Computer Vision and Pattern Recognition* (Providence, RI: IEEE), 2066–2073.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). "Generative adversarial nets," in *Advances in Neural Information Processing Systems* (Montreal, QC), 2672–2680.

Goodfellow, I. J., Erhan, D., Carrier, P., Courville, A., Mirza, M., Hamner, B., et al. (2013). "Challenges in representation learning: a report on three machine learning contests," in *International Conference on Neural Information Processing* (Daegu: Springer), 117–124.

Gopalan, R., Li, R., and Chellappa, R. (2011). "Domain adaptation for object recognition: an unsupervised approach," in *2011 International Conference on Computer Vision* (Barcelona: IEEE), 999–1006.

Grimm, M., Kroschel, K., and Narayanan, S. (2008). "The Vera am Mittag German audio-visual emotional speech database," in *2008 IEEE International Conference on Multimedia and Expo* (Hanover: IEEE), 865–868.

Hansen, J. H., and Bou-Ghazale, S. E. (1997). "Getting started with SUSAS: a speech under simulated and actual stress database," in *Fifth European Conference on Speech Communication and Technology* (Rhodes).

Haq, S., Jackson, P. J., and Edge, J. (2008). "Audio-visual feature selection and reduction for emotion classification," in *Proceedings of the International Conference on Auditory-Visual Speech Processing (AVSP'08)* (Tangalooma, QLD).

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV), 770–778.

Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Trischler, A., and Bengio, Y. (2018). Learning deep representations by mutual information estimation and maximization. *arXiv: 1808.06670*.

Hodges, W., and Spielberger, C. (1966). The effects of threat of shock on heart rate for subjects who differ in manifest anxiety and fear of shock. *Psychophysiology* 2, 287–294. doi: 10.1111/j.1469-8986.1966.tb02656.x

Huang, Y., Yang, J., Liu, S., and Pan, J. (2019). Combining facial expressions and electroencephalography to enhance emotion recognition. *Fut. Int.* 11:105. doi: 10.3390/fi11050105

Jackson, P., and Haq, S. (2014). *Surrey Audio-Visual Expressed Emotion (SAVEE) Database*. Guildford: University of Surrey.

Jiang, Y.-G., Xu, B., and Xue, X. (2014). "Predicting emotions in user-generated videos," in *Twenty-Eighth AAAI Conference on Artificial Intelligence* (Québec, QC).

Jung, H., Ju, J., Jung, M., and Kim, J. (2018). "Less-forgetful learning for domain expansion in deep neural networks," in *Thirty-Second AAAI Conference on Artificial Intelligence* (New Orleans, LA).

Kaya, H., Gürpınar, F., and Salah, A. A. (2017). Video-based emotion recognition in the wild using deep transfer learning and score fusion. *Image Vis. Comput.* 65, 66–75. doi: 10.1016/j.imavis.2017.01.012

Khorram, S., Jaiswal, M., Gideon, J., McInnis, M. G., and Provost, E. M. (2018). The PRIORI emotion dataset: linking mood to emotion detected in-the-wild. *CoRR* abs/1806.10658. doi: 10.21437/Interspeech.2018-2355

Koelstra, S., Muhl, C., Soleymani, M., Lee, J.-S., Yazdani, A., Ebrahimi, T., et al. (2011). DEAP: a database for emotion analysis; using physiological signals. *IEEE Trans. Affect. Comput.* 3, 18–31. doi: 10.1109/T-AFFC.2011.15

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems* (Lake Tahoe, CA), 1097–1105.

Kulis, B., Saenko, K., and Darrell, T. (2011). "What you saw is not what you get: domain adaptation using asymmetric kernel transforms," in *CVPR 2011* (Colorado Springs, CO: IEEE), 1785–1792. doi: 10.1109/CVPR.2011.5995702

Lan, Z., Sourina, O., Wang, L., Scherer, R., and Müller-Putz, G. R. (2018). Domain adaptation techniques for EEG-based emotion recognition: a comparative study on two public datasets. *IEEE Trans. Cogn. Dev. Syst.* 11, 85–94. doi: 10.1109/TCDS.2018.2826840

Latif, S., Qadir, J., and Bilal, M. (2019). "Unsupervised adversarial domain adaptation for cross-lingual speech emotion recognition," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)* (Cambridge: IEEE), 732–737. doi: 10.1109/ACII.2019.8925513

Latif, S., Rana, R., Younis, S., Qadir, J., and Epps, J. (2018). Cross corpus speech emotion classification-an effective transfer learning technique. *arXiv: 1801.06353*.

Li, J., Qiu, S., Shen, Y.-Y., Liu, C.-L., and He, H. (2019). Multisource transfer learning for cross-subject EEG emotion recognition. *IEEE Trans. Cybernet.* doi: 10.1109/TCYB.2019.2904052

Li, Q., and Chaspari, T. (2019). "Exploring transfer learning between scripted and spontaneous speech for emotion recognition," in *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI)* (Suzhou: ACM).

Lin, Y.-P. (2019). Constructing a personalized cross-day EEG-based emotion-classification model using transfer learning. *IEEE J. Biomed. Health Informat.* doi: 10.1109/JBHI.2019.2934172

Lin, Y.-P., Hsu, S.-H., and Jung, T.-P. (2015). "Exploring day-to-day variability in the relations between emotion and eeg signals," in *International Conference on Augmented Cognition* (Los Angeles, CA: Springer), 461–469.

Lin, Y.-P., and Jung, T.-P. (2017). Improving EEG-based emotion classification using conditional transfer learning. *Front. Hum. Neurosci.* 11:334. doi: 10.3389/fnhum.2017.00334

Lin, Y.-P., Wang, C.-H., Jung, T.-P., Wu, T.-L., Jeng, S.-K., Duann, J.-R., et al. (2010). EEG-based emotion recognition in music listening. *IEEE Trans. Biomed. Eng.* 57, 1798–1806. doi: 10.1109/TBME.2010.2048568

Livingstone, S. R., and Russo, F. A. (2018). The ryerson audio-visual database of emotional speech and song (RAVDESS): a dynamic, multimodal set of facial and vocal expressions in north american english. *PLoS ONE* 13:e0196391. doi: 10.1371/journal.pone.0196391

Long, M., Cao, Y., Wang, J., and Jordan, M. I. (2015). Learning transferable features with deep adaptation networks. *arXiv: 1502.02791*.

Long, M., Ding, G., Wang, J., Sun, J., Guo, Y., and Yu, P. S. (2013). "Transfer sparse coding for robust image representation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Portland, OR), 407–414. doi: 10.1109/CVPR.2013.59

Long, M., Zhu, H., Wang, J., and Jordan, M. I. (2017). "Deep transfer learning with joint adaptation networks," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70* (Sydney, NSW), 2208–2217.

Lotfian, R., and Busso, C. (2017). Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Trans. Affect. Comput*.

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., and Matthews, I. (2010). "The extended cohn-kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops* (San Francisco, CA: IEEE), 94–101. doi: 10.1109/CVPRW.2010.5543262

Lyons, M., Akamatsu, S., Kamachi, M., and Gyoba, J. (1998). "Coding facial expressions with gabor wavelets," in *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition* (Nara: IEEE), 200–205.

Lyons, M. J., Budynek, J., and Akamatsu, S. (1999). Automatic classification of single facial images. *IEEE Trans. Patt. Anal. Mach. Intell.* 21, 1357–1362. doi: 10.1109/34.817413

Ma, Y., Hao, Y., Chen, M., Chen, J., Lu, P., and Košir, A. (2019). Audio-visual emotion fusion (avef): a deep efficient weighted approach. *Informat. Fusion* 46, 184–192. doi: 10.1016/j.inffus.2018.06.003

Mao, Q., Xue, W., Rao, Q., Zhang, F., and Zhan, Y. (2016). "Domain adaptation for speech emotion recognition by sharing priors between related source and target classes," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Shanghai: IEEE), 2608–2612.

Martin, O., Kotsia, I., Macq, B., and Pitas, I. (2006). "The eNTERFACE'05 audio-visual emotion database," in *22nd International Conference on Data Engineering Workshops (ICDEW'06)* (Atlanta, GA: IEEE), 8.

Mathias, M., Benenson, R., Pedersoli, M., and Van Gool, L. (2014). "Face detection without bells and whistles," in *European Conference on Computer Vision* (Zurich: Springer), 720–735.

McKeown, G., Valstar, M., Cowie, R., Pantic, M., and Schroder, M. (2011). The semaine database: annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Trans. Affect. Comput.* 3, 5–17. doi: 10.1109/T-AFFC.2011.20

Mollahosseini, A., Hasani, B., and Mahoor, M. H. (2017). Affectnet: a database for facial expression, valence, and arousal computing in the wild. *IEEE Trans. Affect. Comput.* 10, 18–31. doi: 10.1109/TAFFC.2017.2740923

Motiian, S., Piccirilli, M., Adjeroh, D. A., and Doretto, G. (2016). "Information bottleneck learning using privileged information for visual recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV), 1496–1505. doi: 10.1109/CVPR.2016.166

Motiian, S., Piccirilli, M., Adjeroh, D. A., and Doretto, G. (2017). "Unified deep supervised domain adaptation and generalization," in *Proceedings of the IEEE International Conference on Computer Vision* (Venice), 5715–5725. doi: 10.1109/ICCV.2017.609

Muandet, K., Balduzzi, D., and Schölkopf, B. (2013). "Domain generalization via invariant feature representation," in *International Conference on Machine Learning* (Atlanta, GA), 10–18.

Neumann, M., and Vu, N. T. (2019). "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)* (Brighton: IEEE), 7390–7394.

Ng, H.-W., Nguyen, V. D., Vonikakis, V., and Winkler, S. (2015). "Deep learning for emotion recognition on small datasets using transfer learning," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction* (Seattle, WA: ACM), 443–449.

Ngo, T. Q., and Yoon, S. (2019). "Facial expression recognition on static images," in *International Conference on Future Data and Security Engineering* (Nha Trang: Springer), 640–647.

Ortega, J. D., Cardinal, P., and Koerich, A. L. (2019). Emotion recognition using fusion of audio and video features. *arXiv: 1906.10623*. doi: 10.1109/SMC.2019.8914655

Ouyang, X., Kawaai, S., Goh, E. G. H., Shen, S., Ding, W., Ming, H., et al. (2017). "Audio-visual emotion recognition using deep transfer learning and multiple temporal models," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction* (Glasgow: ACM), 577–582.

Pan, S. J., Tsang, I. W., Kwok, J. T., and Yang, Q. (2010). Domain adaptation via transfer component analysis. *IEEE Trans. Neural Netw.* 22, 199–210. doi: 10.1109/TNN.2010.2091281

Pan, S. J., and Yang, Q. (2009). A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* 22, 1345–1359. doi: 10.1109/TKDE.2009.191

Parkhi, O. M., Vedaldi, A., Zisserman, A., et al. (2015). "Deep face recognition," in *BMVC*, Vol. 1 (Swansea), 1–12. doi: 10.5244/C.29.41

Pei, Z., Cao, Z., Long, M., and Wang, J. (2018). "Multi-adversarial domain adaptation," in *Thirty-Second AAAI Conference on Artificial Intelligence* (New Orleans, LA).

PIC (2013). Psychological image collection at stirling (PICS). Available online at: http://pics.stir.ac.uk/

Radford, A., Metz, L., and Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv: 1511.06434*.

Ringeval, F., Sonderegger, A., Sauer, J., and Lalanne, D. (2013). "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)* (Shanghai: IEEE), 1–8. doi: 10.1109/FG.2013.6553805

Ruder, S., and Plank, B. (2017). Learning to select data for transfer learning with bayesian optimization. *arXiv: 1707.05246*. doi: 10.18653/v1/D17-1038

Rusu, A. A., Rabinowitz, N. C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., et al. (2016). Progressive neural networks. *arXiv: 1606.04671*.

Saenko, K., Kulis, B., Fritz, M., and Darrell, T. (2010). "Adapting visual category models to new domains," in *European Conference on Computer Vision* (Heraklion: Springer), 213–226.

Sagha, H., Deng, J., Gavryukova, M., Han, J., and Schuller, B. (2016). "Cross lingual speech emotion recognition using canonical correlation analysis on principal component subspace," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Shanghai: IEEE), 5800–5804.

Sagonas, C., Antonakos, E., Tzimiropoulos, G., Zafeiriou, S., and Pantic, M. (2016). 300 faces in-the-wild challenge: database and results. *Image Vis. Comput.* 47, 3–18. doi: 10.1016/j.imavis.2016.01.002

Schuller, B., Arsic, D., Rigoll, G., Wimmer, M., and Radig, B. (2007). "Audiovisual behavior modeling by combined feature spaces," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, Vol. 2 (Honolulu, HI: IEEE), II–733.

Schuller, B., Müller, R., Eyben, F., Gast, J., Hörnler, B., Wöllmer, M., et al. (2009a). Being bored? Recognising natural interest by extensive audiovisual integration for real-life application. *Image Vis. Comput.* 27, 1760–1774. doi: 10.1016/j.imavis.2009.02.013

Schuller, B., Steidl, S., and Batliner, A. (2009b). "The INTERSPEECH 2009 emotion challenge," in *Tenth Annual Conference of the International Speech Communication Association* (Brighton).

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., et al. (2010). "The interspeech 2010 paralinguistic challenge," in *Eleventh Annual Conference of the International Speech Communication Association* (Makuhari).

Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., et al. (2013). "The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association* (Lyon).

Siegman, A. W., and Boyle, S. (1993). Voices of fear and anxiety and sadness and depression: the effects of speech rate and loudness on fear

and anxiety and sadness and depression. *J. Abnorm. Psychol.* 102:430. doi: 10.1037/0021-843X.102.3.430

Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv: 1409.1556*.

Song, P. (2017). Transfer linear subspace learning for cross-corpus speech emotion recognition. *IEEE Trans. Affect. Comput.* doi: 10.1109/TAFFC.2018.2800046

Song, P., Zheng, W., Liu, J., Li, J., and Zhang, X. (2015). "A novel speech emotion recognition method via transfer pca and sparse coding," in *Chinese Conference on Biometric Recognition* (Urumchi: Springer), 393–400.

Staroniewicz, P., and Majewski, W. (2009). "Polish emotional speech database–recording and preliminary validation," in *Cross-Modal Analysis of Speech, Gestures, Gaze and Facial Expressions* (Prague: Springer), 42–49.

Steidl, S. (2009). *Automatic Classification of Emotion Related User States in Spontaneous Children's Speech*. University of Erlangen-Nuremberg Erlangen.

Sugianto, N., and Tjondronegoro, D. (2019). "Cross-domain knowledge transfer for incremental deep learning in facial expression recognition," in *2019 7th International Conference on Robot Intelligence Technology and Applications (RiTA)* (Daejeon: IEEE), 205–209.

Sun, S., Yeh, C.-F., Hwang, M.-Y., Ostendorf, M., and Xie, L. (2018). "Domain adversarial training for accented speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Calgary, AB: IEEE), 4854–4858.

Tommasi, T., Lanzi, M., Russo, P., and Caputo, B. (2016). "Learning the roots of visual domain shift," in *European Conference on Computer Vision* (Amsterdam: Springer), 475–482.

Tzeng, E., Hoffman, J., Darrell, T., and Saenko, K. (2015). "Simultaneous deep transfer across domains and tasks," in *Proceedings of the IEEE International Conference on Computer Vision* (Santiago), 4068–4076.

Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., and Darrell, T. (2014). Deep domain confusion: maximizing for domain invariance. *arXiv: 1412.3474*.

Valstar, M., and Pantic, M. (2010). "Induced disgust, happiness and surprise: an addition to the mmi facial expression database," in *Proceedings of the 3rd International Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect* (Paris), 65.

Vielzeuf, V., Pateux, S., and Jurie, F. (2017). "Temporal multimodal fusion for video emotion classification in the wild," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction* (Glasgow: ACM), 569–576.

Wang, D., and Zheng, T. F. (2015). "Transfer learning for speech and language processing," in *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)* (Hong Kong: IEEE), 1225–1237.

Wöllmer, M., Metallinou, A., Eyben, F., Schuller, B., and Narayanan, S. (2010). "Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling," in *Proceedings of the INTERSPEECH 2010* (Makuhari), 2362–2365.

Xu, B., Fu, Y., Jiang, Y.-G., Li, B., and Sigal, L. (2016). "Video emotion recognition with transferred deep feature encodings," in *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval* (New York, NY: ACM), 15–22.

Xu, M., Cheng, W., Zhao, Q., Ma, L., and Xu, F. (2015). "Facial expression recognition based on transfer learning from deep convolutional networks," in *2015 11th International Conference on Natural Computation (ICNC)* (Zhangjiajie: IEEE), 702–708.

Xu, R., Chen, Z., Zuo, W., Yan, J., and Lin, L. (2018). "Deep cocktail network: multi-source unsupervised domain adaptation with category shift," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT), 3964–3973.

Yan, J., Zheng, W., Cui, Z., Tang, C., Zhang, T., and Zong, Y. (2018). Multi-cue fusion for emotion recognition in the wild. *Neurocomputing* 309, 27–35. doi: 10.1016/j.neucom.2018.03.068

Yang, J., Yan, R., and Hauptmann, A. G. (2007). "Adapting SVM classifiers to data with shifted distributions," in *Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007)* (Omaha, NE: IEEE), 69–76.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). "How transferable are features in deep neural networks?," in *Advances in Neural Information Processing Systems* (Montreal, QC), 3320–3328.

Zhang, B., Provost, E. M., and Essl, G. (2016). "Cross-corpus acoustic emotion recognition from singing and speaking: a multi-task learning approach," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Shanghai, IEEE), 5805–5809.

Zhang, B., Provost, E. M., Swedberg, R., and Essl, G. (2015). "Predicting emotion perception across domains: a study of singing and speaking," in *Twenty-Ninth AAAI Conference on Artificial Intelligence* (Austin, TX).

Zhang, S., Zhang, S., Huang, T., Gao, W., and Tian, Q. (2017). Learning affective features with a hybrid deep model for audio–visual emotion recognition. *IEEE Trans. Circ. Syst. Video Technol.* 28, 3030–3043. doi: 10.1109/TCSVT.2017.2719043

Zhang, W., Wang, F., Jiang, Y., Xu, Z., Wu, S., and Zhang, Y. (2019). "Cross-subject EEG-based emotion recognition with deep domain confusion," in *International Conference on Intelligent Robotics and Applications* (Shenyang: Springer), 558–570.

Zhang, X., Zhang, L., Wang, X.-J., and Shum, H.-Y. (2012). Finding celebrities in billions of web images. *IEEE Trans. Multimedia* 14, 995–1007. doi: 10.1109/TMM.2012.2186121

Zhao, H., Ye, N., and Wang, R. (2019). Speech emotion recognition based on hierarchical attributes using feature nets. *Int. J. Parallel Emergent Distrib. Syst.* 1–11. doi: 10.1080/17445760.2019.1626854

Zhao, H., Zhang, S., Wu, G., Moura, J. M. F., Costeira, J. P., and Gordon, G. J. (2018). "Adversarial multiple source domain adaptation," in *Advances in Neural Information Processing Systems 31*, eds S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Montreal, QC: Curran Associates, Inc.), 8559–8570.

Zheng, W.-L., and Lu, B.-L. (2015). Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. *IEEE Trans. Auton. Ment. Dev.* 7, 162–175.

Zheng, W.-L., and Lu, B.-L. (2016). "Personalizing eeg-based affective models with transfer learning," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence* (New York, NY: AAAI Press), 2732–2738.

Zheng, W.-L., Zhang, Y.-Q., Zhu, J.-Y., and Lu, B.-L. (2015). "Transfer components between subjects for EEG-based emotion recognition," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)* (Xi'an: IEEE), 917–922.

Zhou, H., and Chen, K. (2019). "Transferable positive/negative speech emotion recognition via class-wise adversarial domain adaptation," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Brighton, VIC: IEEE), 3732–3736.

Zong, Y., Zheng, W., Zhang, T., and Huang, X. (2016). Cross-corpus speech emotion recognition based on domain-adaptive least-squares regression. *IEEE Signal Process. Lett.* 23, 585–589. doi: 10.1109/LSP.2016.2537926

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.