IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 27, 2021, accepted June 27, 2021, date of publication July 2, 2021, date of current version July 13, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3094355

Transfer Subspace Learning for Unsupervised Cross-Corpus Speech Emotion Recognition

NA LIU[1], BAOFENG ZHANG[1], BIN LIU[1], JINGANG SHI[2], (Member, IEEE), LEI YANG[3], ZHIWEI LI[1], AND JUNCHAO ZHU[1]

[1]Tianjin Key Laboratory for Control Theory and Applications in Complicated System, Tianjin University of Technology, Tianjin 300384, China
[2]School of Software, Xi'an Jiaotong University, Xi'an 710000, China
[3]Autobrain (Tianjin) Technology Company Ltd., Tianjin 300300, China

Corresponding author: Junchao Zhu (zhujunchao_tjut@163.com)

This work was supported in part by the Key Training Project for Tianjin Project Plus Team under Grant XC202054, in part by the Tianjin Science and Technology Project under Grant 18YFCZZC00320, in part by the National Natural Science Foundation of China under Grant 61172185 and Grant 62002283, and in part by the Application Foundation and Advanced Technology Research Project of Tianjin.

ABSTRACT In many practical applications, a speech emotion recognition model learned on a source (training) domain but applied to a novel target (testing) domain degenerates even significantly due to the mismatch between the two domains. Aiming at learning a better speech emotion recognition model for the target domain, the paper investigates this interesting problem, i.e., unsupervised cross-corpus speech emotion recognition (SER), in which the training and testing speech signals come from two different speech emotion corpora. Meanwhile, the training speech signals are labeled, while the label information of the testing speech signals is entirely unknown. To deal with this problem, we propose a simple yet effective method called transfer subspace learning (TRaSL). TRaSL aims at learning a projection matrix with which we can transform the source and target speech signals from the original feature space to the label space. The transformed source and target speech signals in the label space would share similar feature distributions. Consequently, the classifier learned on the labeled source speech signals can effectively predict the emotional states of the unlabeled target speech signals. To evaluate the performance of the proposed TRaSL method, we carry out extensive cross-corpus SER experiments on four speech emotion corpora including IEMOCAP, EmoDB, eNTERFACE, and AFEW 4.0. Compared with recent state-of-the-art cross-corpus SER methods, the proposed TRaSL can achieve more satisfactory overall results.

INDEX TERMS Cross-corpus speech emotion recognition, subspace learning, transfer learning, domain adaptation.

I. INTRODUCTION

Speech emotion recognition (SER) has been a very attractive research field in affective computing, pattern recognition, and human-computer interaction (HCI). A major task of speech emotion recognition is to provide computers the ability to recognize the human beings' emotional states such as happy, angry, and disgust from their speech signals [1]. In recent years, extensive effective methods have been proposed to deal with this problem [2], [3]. But it can be noted that most of the current speech emotion recognition methods are heavily dependent on one common assumption, namely that the training speech samples and the testing one belong to the same corpus. In this case, it can be thought that the speech signals extracted from training and testing speech sequences abide by the same or similar marginal probability distribution. In many practical situations, however, the training and testing samples may belong to different domains, e.g., the training speech samples and the testing ones are recorded by different equipment or collected under different environments. Hence in this scenario, the marginal probability distribution of emotion signal vector set in training speech would quite different from that in testing ones. This thus creates a more difficult yet interesting problem than conventional SER, i.e., cross-corpus SER. To distinguish the training and testing speech corpora in cross-corpus SER problem, these two corpora can be referred as source corpus and target corpus, respectively. In the work of [4], Deng *et al.* classified cross- corpus speech emotion recognition into two categories including semi-supervised case and unsupervised case. The main difference between

The associate editor coordinating the review of this manuscript and approving it for publication was Mohammad Zia Ur Rahman.

these two categories is whether we can get the label information of target domain. Homogeneously, cross-corpus SER can follow this classification. In this paper, we will investigate the unsupervised cross-corpus SER, in which the training and testing speech signals come from two different speech emotion corpora. Meanwhile, the training speech signals are labeled, while the label information of the testing speech signals is completely unknown. Due to this setting, the training and testing speech signals may have different feature distributions. To deal with this problem, we propose a novel method called transfer subspace learning (TRaSL). Our preliminary work [5] reduced the discrepancy of the source and domain to complete the classification, but the structures of both domains were not been considered. The basic idea of TRaSL is to learn a projection matrix which transforms the source and target speech signals from the original feature space to a common subspace. In such common space, the source and target speech signals are enforced to obey the similar feature distributions and hence we can train a classifier, e.g., support vector machine (SVM), based on the labeled source speech signals such that it can accurately predict the emotional states of the target speech signals. Motivated by the works of [6], [7], we construct a label space based on the label information provided by the source speech corpora to serve as the predefined common subspace for TRaSL.

The main contributions of this paper for unsupervised cross-corpus speech emotion recognition are summarized as follows:

1) A new framework called TRaSL for dealing with unsupervised cross-corpus speech emotion recognition is proposed. In the TRaSL model:
   (a) A projection matrix is learnt to transform the source and target speech signals from the original feature space to the common space.
   (b) In the common space the disparity of source and target feature vectors is reduced.
   (c) The structures of source and target domains are enforced to be approximation, which can keep enough discriminant information for further model learning.
2) We use four representative cross-corpus SER methods and SVM as baseline to conduct more extensive evaluation experiments under the designed protocol and deeply discuss the experimental results.

The remainder of this paper is organized as follows: Section II presents recent works about cross-corpus SER. In Section III, we describe the central idea of TRaSL framework for cross-corpus SER, along with the optimization method to solve this issue. For evaluating our TRaSL framework, extensive experiments are conducted in Section IV. Finally, we conclude our paper in Section V.

II. RELATED WORKS

A. DOMAIN ADAPTATION

Domain adaptation (DA) is a representative method in transfer learning, which uses labeled source domain samples to improve the performance of target domain model [20]. DA problem is that labeled source domain and unlabeled target domain share the same categories, but the distribution of features is different, i.e., $X^s \neq X^t$: $P^s(X) \neq P^t(X)$, where X^s and X^t are the feature matrices, $P^s(X)$ and $P^t(X)$ are the feature distributions of source and target domain, respectively. DA can be broadly categorized into two groups according to whether the target domain sample has some labels or it is entirely unlabeled. The former is referred to as semi-supervised DA, while the latter is called unsupervised DA. While semi-supervised DA is generally performed by utilizing the correspondence information obtained from labeled target domain data to learn the domain shifting transformation (e.g. [6]), unsupervised DA is based on the following strategies: (i) imposing certain assumptions on the class of transformations between domains [8], or (ii) assuming the availability of certain discriminative features that are common to both domains [9], [10], [13].

B. CROSS-CORPUS SPEECH EMOTION RECOGNITION

Cross-corpus SER is a new learning setting which allows source and target samples to come from different distributions. Consequently, how to deal with this problem is an important and challenging case in current research. In spite of that, some researchers have focused on this challenging problem and proposed some effective methods. Schuller *et al.* [11] attempted to employ multiple normalization schemes to investigate cross-corpus SER problem, which may be the first research about cross-corpus SER. Thereafter, more diverse cross-corpus SER methods are in sequence proposed [4], [6], [12], [14]–[18]. For example, Deng *et al.* [4], [12], [14] proposed a series of autoencoder based domain adaptation methods to deal with cross-corpus SER, in which autoencoder networks are exploited to learn the new representations for source and target speech samples. In the work of [15], Hassan *et al.* proposed an importance weighted support vector machine (IW-SVM) to cope with cross-corpus SER problems. IW-SVM leverages three transfer learning methods [20], i.e., kernel mean matching (KMM) [21], Kullback-Leibler importance estimation procedure (KLIEP) [9], and unconstrained least squares importance fitting (uLSIF) [23], to learn a set of importance weights for target speech samples such that the feature distribution mismatch between source and target speech samples is relieved. Besides the above methods, it is also worth mentioning the work of selective transfer machine (STM) [24], [25], which is proposed for personalized (cross-subject) facial action unit detection. STM inherits the ability of kernel mean matching (KMM) [21] to eliminate the feature distribution difference between source and target samples and also have the discriminative ability of support vector machine (SVM). The abovementioned subspace learning algorithms focus on finding the latent common feature representations to cope with the feature matching problem, and do not take into account the importance of feature selection together.

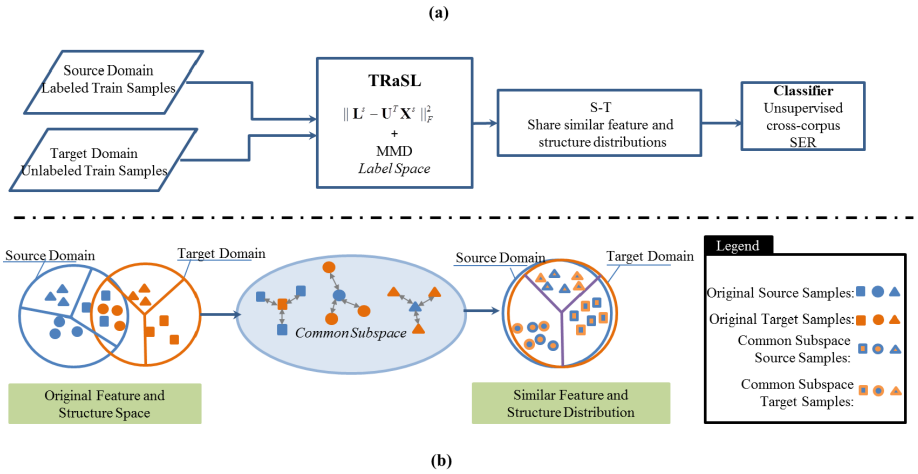

FIGURE 1. The overall schema of the proposed method: (a) the overall pipeline of our domain adaptation method: first to learn a projection matrix which transforms the source and target speech signals from the original feature space to the label space, and then minimize distance difference between the source and target domain; (b) the projection matrix used in (a) are learnt in a common subspace rather than in the original sample space, where the source and target speech features will be enforced to share the similar distribution.

Recently, a transfer non-negative matrix factorization (TNNMF) method is proposed by Song *et al.* [16] for cross-corpus SER tasks. In TNNMF, the maximum mean discrepancy (MMD) [26] is used to balance the feature distribution difference between the originally distinct source and target speech signals. Zong *et al.* [6], [7] proposed a novel domain adaptation method called domain adaptive least squares regression (DALSR) model to handle cross-corpus SER. DALSR aims at learning a regression coefficient matrix to bridge the source and target speech corpora. Though DALSR considered the importance of feature selection, it should fix the number of auxiliary samples of the target corpus. More recently, Song *et al.* [40] also presented a feature selection based transfer subspace learning (FSTSL) method to cope with cross-corpus SER problem, which considers feature selection as an additional constraint. FSTSL considered the feature distribution difference, while neglected the discriminant information of the model.

Besides these studies, a label space is constructed according to the label information provided by the source speech corpora, which serves as the predefined common subspace for TRaSL. TRaSL considered the feature distribution difference and the discriminant property of the model, meanwhile, importance of feature selection also take into account.

III. PROPOSED METHOD

A. BASIC IDEA OF TRaSL

In this section, we firstly introduce the basic idea of TRaSL. For better understanding of TRaSL framework, Fig. 1 gives the architecture of proposed model. It can be seen from Fig. 1(a) that the goal of the proposed TRaSL framework is to learn a projection matrix with which the source and target speech signals can be transformed from the original feature space to the label space. In the label space the source and target speech signals would share similar feature and structure distribution which is depicted in Fig. 1(b). What follows is to train a classifier. Using the projected source speech features and its given label information to predict the projected target signal categories.

B. TRaSL MODEL FRAMEWORK

For clarity, we define some notations. In the whole text, matrices are written in upper-case letters, vectors are written as lower-case letters.

Suppose we have two different speech corpora to serve as source and target corpus, respectively. Their corresponding feature matrices are denoted by $\mathbf{X}^s \in \mathbf{R}^{d \times N_s}$ and $\mathbf{X}^t \in \mathbf{R}^{d \times N_t}$, where d is the dimension of the speech feature vector and N_s and N_t are the numbers of the source and target speech signals, respectively. For unsupervised cross-corpus SER case, the label information of source speech signals is available, thus we denote their label information as the vector form, which is followed by the works of [4], [6]. Specifically, let $\mathbf{L}^s \in \mathbf{R}^{c \times N_s}$ be the label matrix corresponding to the source feature matrix $\mathbf{X}^s$, where c is the number of speech emotion states, and the i^{th} column $\mathbf{1}_i^s = [l_{i,1}^s, \cdots, l_{i,c}^s]^T$ is a class label vector of $\mathbf{L}^s$ whose elements will take the value of 0 or 1 according to the following rule:

$$l_{i,j} = \begin{cases} 1, & \text{if } \mathrm{x}_i^s \text{ belongs to the } j^{th} \text{ emotion states;} \\ 0, & \text{otherwise} \end{cases}$$

By using these source label vectors, we are thus able to construct a new subspace as the predefined common subspace. Note that our TRaSL aims at learning a projection matrix $\mathbf{U}$ to project the source speech feature matrix $\mathbf{X}^s$ from the original feature space to such common subspace spanned by the columns of $\mathbf{L}^s$, which can be formulated as the

following optimization problem:

$$\min_{\mathbf{U}} ||\mathbf{L}^s - \mathbf{U}^T\mathbf{X}^s||_F^2 \tag{1}$$

Meanwhile, with the projection matrix $\mathbf{U}$, the target speech feature matrix $\mathbf{X}^t$ can also be mapped to the predefined common subspace, where the projected source and target speech features will be enforced to share the similar distributions. To achieve this goal, following the works of MMD criterion [26] and TNNMF [16], we minimize the distance difference between mean projected source speech feature vectors and mean projected target speech feature vectors, which are formulated as follows:

$$\min_{\mathbf{U}} ||\frac{1}{N_s}\sum_{i=1}^{N_s}\mathbf{U}^T\mathbf{x}_i^s - \frac{1}{N_t}\sum_{i=1}^{N_t}\mathbf{U}^T\mathbf{x}_i^t||^2 \tag{2}$$

Besides reducing the discrepancy of the source and target domains, in the common subspace the structures of both domains are also expected to be the same. Motivated by the work of Gretton *et al.* [28], we propose to impose the projected covariance matrix difference between the source speech feature vectors and the target ones in the subspace on to the objective of Eqs.(1) and (2), the structure is limited just by simply minimizing the variance of each domain, which can be formulated as Eq. (3):

$$||\mathbf{U}^T[\frac{1}{N_s}\sum_{i=1}^{N_s}(\mathbf{x}_i^s - \bar{x}^s)(\mathbf{x}_i^s - \bar{x}^s)^T - \frac{1}{N_t}\sum_{i=1}^{N_t}(\mathbf{x}_i^t - \bar{x}^t)(\mathbf{x}_i^t - \bar{x}^t)^T]\mathbf{U}||_F^2 \tag{3}$$

with $\bar{x}^s = 1/N_s\sum_{i=1}^{N_s}\mathbf{x}_i^s, \bar{x}^t = 1/N_t\sum_{i=1}^{N_t}\mathbf{x}_i^t$.

By minimizing the combination of the above objective functions in Eqs. (1), (2) and (3), we can arrive at the final optimization problem as the following formulation:

$$\min_{\mathbf{U}} ||\mathbf{L}^s - \mathbf{U}^T\mathbf{X}^s||_F^2 + \lambda_2||\mathbf{U}^T||_{2,1} + \lambda_1(||\mathbf{U}^T\bar{x}^s - \mathbf{U}^T\bar{x}^t||^2 + ||\mathbf{U}^T(\Sigma^s - \Sigma^t)\mathbf{U}||_F^2) \tag{4}$$

where λ_1 and λ_2 are the trade-off parameters to control the balance among three terms in the objective functions. $\Sigma^s = 1/N_s\Sigma_{i=1}^{N_s}(\mathbf{x}_i^s - \bar{\mathbf{x}}^s)(\mathbf{x}_i^s - \bar{\mathbf{x}}^s)^T$ and $\Sigma^t = 1/N_t\Sigma_{i=1}^{N_t}(\mathbf{x}_i^t - \bar{\mathbf{x}}^t)(\mathbf{x}_i^t - \bar{\mathbf{x}}^t)^T$. It should be also noted that besides previously described combination, we introduce a $L_{2,1}$ norm term with respect to the transpose matrix of $\mathbf{U}$ to serve as the regularization to select the important features contributing to SER [6] during the feature projection. Then we can get our TRaSL model, which is shown in Eq. (4).

C. OPTIMIZATION OF DOSL FRAMEWORK

TRaSL model can be solved by using inexact augmented Lagrange multiplier (IALM) method [27]. More specifically, by introducing two auxiliary variables $\mathbf{Q}$ and $\mathbf{K}$, which satisfies $\mathbf{U} = \mathbf{Q}$, $\mathbf{U} = \mathbf{K}$, first Eq. (4) can be reformulated as:

$$\min_{\mathbf{U}} ||\mathbf{L}^s - \mathbf{U}^T\mathbf{X}^s||_F^2 + \lambda_1(||\mathbf{U}^T\Delta\bar{\mathbf{x}}^{st}||^2 + ||\mathbf{U}^T\Delta\Sigma^{st}\mathbf{U}||_F^2) + \lambda_2||\mathbf{U}^T||_{2,1} \tag{5}$$

with $\Delta\bar{\mathbf{x}}^{st} = \bar{\mathbf{x}}^s - \bar{\mathbf{x}}^t$, $\Delta\Sigma^{st} = \Sigma^s - \Sigma^t$
we convert the optimization problem of (5) to a constrained one which can be expressed as:

$$\begin{aligned}&\min_{\mathbf{U},\mathbf{K},\mathbf{Q}} ||\mathbf{L}^s - \mathbf{K}^T\mathbf{X}^s||_F^2 \\ &\quad +\lambda_1(||\mathbf{K}^T\Delta\bar{\mathbf{x}}^{st}||^2 + ||\mathbf{Q}^T\Delta\Sigma^{st}\mathbf{K}||_F^2) + \lambda_2||\mathbf{U}^T||_{2,1} \\ &s.t. \quad \mathbf{U} = \mathbf{K} \text{ and } \mathbf{U} = \mathbf{Q}\end{aligned} \tag{6}$$

Subsequently, the Lagrange function of Eq. (6) can be obtained as follows:

$$\begin{aligned}&L(\mathbf{U}, \mathbf{K}, \mathbf{Q}, \mathbf{T}_1, \mathbf{T}_2, \mu) \\ &= ||\mathbf{L}^s - \mathbf{K}^T\mathbf{X}^s||_F^2 \\ &\quad +\lambda_1(||\mathbf{K}^T\Delta\bar{\mathbf{x}}^{st}||^2 + ||\mathbf{Q}^T\Delta\Sigma^{st}\mathbf{K}||_F^2) + \lambda_2||\mathbf{U}^T||_{2,1} \\ &\quad +tr[\mathbf{T}_1^T(\mathbf{U} - \mathbf{K})] + tr[\mathbf{T}_2^T(\mathbf{U} - \mathbf{Q})] + \frac{\mu}{2}(||\mathbf{U} - \mathbf{K}||_F^2 \\ &\quad +||\mathbf{U} - \mathbf{Q}||_F^2)\end{aligned} \tag{7}$$

where $\mathbf{T}_1$ and $\mathbf{T}_2$ are the Lagrange multiplier, and $\mu > 0$ is the regularization parameter.

Finally, to achieve the optimal solution of $\mathbf{U}$, we only need to iteratively minimize the Lagrange function of Eq. (7) with respect to one of the variables fixing the others until convergence. More specifically, perform the following five steps:

1. Fix $\mathbf{U}$, $\mathbf{Q}$, $\mathbf{T}_1$, $\mathbf{T}_2$ and μ, update $\mathbf{K}$: In this case, the optimization problem would become as below:

$$\min_{\mathbf{K}} ||\mathbf{L}^s - \mathbf{K}^T\mathbf{X}^s||_F^2 + \lambda_1(||\mathbf{K}^T\Delta\bar{\mathbf{x}}^{st}||^2 + ||\mathbf{Q}^T\Delta\Sigma^{st}\mathbf{K}||_F^2) + tr[\mathbf{T}_1^T(\mathbf{U} - \mathbf{K})] + \frac{\mu}{2}||\mathbf{U} - \mathbf{K}||_F^2 0$$

which results in $\mathbf{K}$, as shown at the bottom of the page, where $\mathbf{I}$ is the identity matrix.

$$\mathbf{K} = (2 \times \frac{\mathbf{X}^s\mathbf{X}^{sT} + \lambda_1\left(\Delta\bar{\mathbf{x}}^{-st}\Delta\bar{\mathbf{x}}^{-stT} + \Delta\Sigma^{st}\mathbf{Q}\mathbf{Q}^T\Delta\Sigma^{st}\right)}{\mu} + \mathbf{I}\Bigg)^{-1} \times \left(\frac{2\mathbf{X}^s\mathbf{L}^{sT} - \mathbf{T}_1}{\mu} + \mathbf{U}\right)$$

2. Fix $\mathbf{U}$, $\mathbf{K}$, $\mathbf{T}_1$, $\mathbf{T}_2$ and μ, update $\mathbf{Q}$: The Eq. (7) can be obtained as follows:

$$\min_{\mathbf{Q}} \lambda_1 ||\mathbf{Q}^T \Delta \Sigma^{st} \mathbf{K}||_F^2 + \frac{\mu}{2} ||\mathbf{U} - \mathbf{Q}||_F^2 || + tr[\mathbf{T}_2^T (\mathbf{U} - \mathbf{Q})]$$

$$\mathbf{Q} = (\frac{2\lambda_1 \Delta \Sigma^{st} \mathbf{K} \mathbf{K}^T \Delta \Sigma^{st})}{\mu} + \mathbf{I})^{-1} (\mathbf{U} - \frac{\mathbf{T}_1^T + \mathbf{T}_2^T}{\mu})$$

3. Fix $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{T}_1$, $\mathbf{T}_2$ and μ, update $\mathbf{U}$: The optimization problem can be rewritten as the following formulation:

$$\min_{\mathbf{U}} \frac{\lambda_2}{\mu} ||\mathbf{U}^T||_{2,1} + \frac{1}{2} ||\mathbf{U}^T - (\frac{\mathbf{Q}^T + \mathbf{K}^T}{2} + \frac{\mathbf{T}_1^T + \mathbf{T}_2^T}{2\mu})||_F^2$$

According to Lemma 4.1 in [15], the optimal $\mathbf{U}$ can be obtained as follows:
if $\frac{\lambda_2}{2\mu} < ||\frac{\mathbf{q}_i + \mathbf{k}_i}{2} + \frac{\mathbf{t}_{1i} + \mathbf{t}_{2i}}{2\mu}||$,

$$\mathbf{u}_i = \frac{||\frac{\mathbf{q}_i + \mathbf{k}_i}{2} + \frac{\mathbf{t}_{1i} + \mathbf{t}_{2i}}{2\mu}|| - \frac{\lambda_2}{2\mu}}{||\frac{\mathbf{q}_i + \mathbf{k}_i}{2} + \frac{\mathbf{t}_{1i} + \mathbf{t}_{2i}}{2\mu}||} (\frac{\mathbf{q}_i + \mathbf{k}_i}{2} + \frac{\mathbf{t}_{1i} + \mathbf{t}_{2i}}{2\mu})$$

otherwise $\mathbf{u}_i = 0$, where $\mathbf{q}_i$, $\mathbf{t}_{1i}$, $\mathbf{t}_{2i}$ and $\mathbf{k}_i$ are the i^{th} row of $\mathbf{Q}, \mathbf{T}_1, \mathbf{T}_2$ and $\mathbf{K}$, respectively.

4. Update $\mathbf{T}_1$, $\mathbf{T}_2$ and μ:

$$\mathbf{T}_1 = \mathbf{T}_1 + \mu(\mathbf{U} - \mathbf{K}), \mathbf{T}_2 = \mathbf{T}_2 + \mu(\mathbf{U} - \mathbf{Q})$$
$$\mu = \max(\mu_{\max}, \rho\mu)$$

where ρ is a scale parameter.

5. Check convergence: $||\mathbf{U} - \mathbf{K}||_\infty < \varepsilon$, $||\mathbf{U} - \mathbf{Q}||_\infty < \varepsilon$ where ε denotes the machine epsilon.

D. CROSS-CORPUS SER USING TRaSL MODEL

By using the above solving method in Section 3.2 to learn the optimal $\mathbf{U}_*$, we have following method to predict the emotion states of the target speech samples. It is to assign the emotion labels to the target speech signals according to the criterion: $emotion\_labels = \arg\max_k \{[\mathbf{U}_*^T \mathbf{X}^t](:, k)\}$, where $[\mathbf{U}_*^T \mathbf{X}^t](:, k)$ means the k^{th} element of the j^{th} column (target speech signal) of the projected matrix $\mathbf{U}_*^T \mathbf{X}^t$.

IV. EXPERIMENTS AND DISCUSSION

A. SPEECH EMOTION DATABASE

In this section, we conduct extensive cross-corpus SER experiments to evaluate the performance of the proposed TRaSL method. Four popular speech emotion corpora including EmoDB [30], the audio dataset of eNTERFACE [31], the audio dataset of AFEW 4.0 [32], and the IEMOCAP database [34] are employed. The detailed information of the above speech emotion database is shown in Table 1.

- The first dataset is **EmoDB**. It covers seven emotion categories: Happiness, Anger, Disgust, Fear, Sadness, Neutral and Surprise. 10 (5 female) professional actors speak 10 German emotionally undefined sentences, including 535 samples.
- The second dataset is **eNTERFACE** corpus. eNTERFACE is composed of 1287 emotion videos from 43 subjects and they are categorized into six basic emotions including Happy, Anger, Disgust, Fear, Sadness and Surprise.
- The third dataset is **AFEW 4.0**. This dataset includes three subsets: Train (578 samples), Val (383 samples) and Test (407 samples).
- The fourth dataset is **IEMOCAP**. It provided by the University of Southern California (USC) which consists often speakers (5 male and 5 female) with five sessions, each session recorded with one male and one female speaker. The database provided 10 emotion categories, i.e., Happy, Angry, Disgust, Neutral, Sad, Fear, Surprise, Frustration, Excited and Other. For the experiments, we only selected the utterances with agreement between 74.6% annotators which lead to 5531 utterances from the four emotions consist of neutral (1708), angry (1103), sad (1084) and happy (1636). The happy class includes both happy and excitement classes. This is the standard data selection used in many experiments using IEM-OCAP database [10], [36], [37], [38].

TABLE 1. The sample numbers of EmoDB, eNTERFACE, AFEW4.0, IEMOCAP database for cross-corpus speech emotion recognition experiments.

Database	Speech Emotion Category						
	Surprise	Angry	Happy	Fear	Sad	Neutral	Disgust
EmoDB	—	127	71	69	62	79	81
eNTERFACE	215	215	212	215	215	—	215
AFEW4.0	103	156	171	113	145	167	106
IEMOCAP	—	1103	1636	—	1084	1708	—

B. EXPERIMENTAL SETTINGS

Following the experimental protocol of [5], [6], we select any two datasets of speech corpora each time and select the samples belonging to the common emotion states from these two datasets, which are served as source and target corpus, alternatively. Therefore, there are finally twelve experiments and each group of experiment consists of two sub-experiments. For convenience, these twelve experiments are denoted by Exp.1, Exp.2, $\cdots$, Exp.12, respectively, whose detailed source and target speech corpora are illustrated in Table 2 and Table 3. Additionally, Figure2 gives a detailed description of common emotion states used in each group, e.g., in Exp.1 and Exp.2, the two employed datasets in this combination are alternatively served as source and target databases, where the common emotion states are Angry, Happy, Sad and Neutral. In Exp.3 and Exp.4, the common emotion states are Angry, Happy and Sad. The vertical axis represents source database samples number in the experiments.

We utilize the INTERSPEECH 2009 feature set as the speech emotion features which can be extracted with the open source OpenSMILE software [35]. The speech signal consists of 384 elements, i.e., 16 acoustic low-level descriptors (LLDs), such as zero-crossing-rate (ZCR), root mean square frame energy (RMS Energy), Mel-frequency cepstral

TABLE 2. Results of the exp.1 to exp.12 cross-corpus speech emotion recognition experiments in terms of WAR, we select the common emotion states for the comparative experiment of each group, in which the best results are highlighted in bold.

Exp.	Source Corpus	Target Corpus	Method							
			SVM [8]	KMM [21]	KLIEP [9]	uLSIF [23]	DALSR[6]	DoSL[5]	TRaSL	Avg
1	IEMOCAP	EmoDB	41.30	37.17	37.46	37.76	66.08	**67.85**	65.49	**50.44**
2	EmoDB	IEMOCAP	33.54	19.60	23.56	43.55	43.57	**45.09**	44.01	36.13
3	IEMOCAP	eNTERFACE	**47.29**	33.33	33.02	36.43	43.88	46.51	45.43	40.84
4	eNTERFACE	IEMOCAP	39.79	28.85	34.61	41.88	51.69	**55.35**	55.30	43.92
5	IEMOCAP	AFEW4.0	31.61	26.76	26.76	38.81	39.75	**40.69**	**40.69**	35.01
6	AFEW4.0	IEMOCAP	34.71	31.15	23.61	40.12	40.68	**41.53**	40.90	36.10
7	EmoDB	eNTERFACE	30.08	23.14	21.82	25.75	36.40	37.51	**37.79**	30.40
8	eNTERFACE	EmoDB	24.27	44.69	27.01	42.27	**52.27**	52.00	**52.27**	42.11
9	EmoDB	AFEW4.0	25.99	29.78	25.57	25.93	30.19	31.00	**31.24**	28.53
10	AFEW4.0	EmoDB	35.02	46.81	31.37	44.38	47.80	50.00	**50.33**	43.67
11	eNTERFACE	AFEW4.0	18.39	25.72	19.60	21.21	26.70	26.20	**27.30**	23.59
12	AFEW4.0	eNTERFACE	18.72	19.75	17.47	18.11	21.96	21.66	**22.27**	19.99
Avg	-	-	31.73	30.56	26.82	34.68	41.75	**42.95**	42.75	35.89

TABLE 3. Results of the exp.1 to exp.12 cross-corpus speech emotion recognition experiments in terms of UAR, we select the common emotion states for the comparative experiment of each group, in which the best results are highlighted in bold.

Exp.	Source Corpus	Target Corpus	Method							
			SVM [8]	KMM [21]	KLIEP [9]	uLSIF [23]	DALSR[6]	DoSL[5]	TRaSL	Avg
1	IEMOCAP	EmoDB	29.81	24.80	25.00	25.23	62.32	**62.91**	61.02	41.58
2	EmoDB	IEMOCAP	36.05	25.00	28.81	43.38	43.43	46.80	**47.90**	38.77
3	IEMOCAP	eNTERFACE	**47.22**	33.33	33.49	36.63	43.90	46.54	45.41	40.93
4	eNTERFACE	IEMOCAP	46.27	33.33	40.03	40.25	54.10	56.03	**56.52**	**46.65**
5	IEMOCAP	AFEW4.0	32.63	25.00	21.78	39.13	40.29	**40.36**	**40.36**	34.22
6	AFEW4.0	IEMOCAP	32.84	25.51	27.38	40.83	41.12	41.42	**42.08**	35.88
7	EmoDB	eNTERFACE	30.06	23.08	21.79	25.75	36.36	37.49	**37.76**	29.14
8	eNTERFACE	EmoDB	27.83	40.18	28.58	40.42	**44.41**	44.25	43.44	37.61
9	EmoDB	AFEW4.0	26.07	**30.39**	25.47	25.75	27.51	29.10	29.38	27.38
10	AFEW4.0	EmoDB	29.87	38.17	27.41	36.25	37.33	39.66	**39.93**	34.78
11	eNTERFACE	AFEW4.0	20.80	23.79	18.66	22.61	24.67	24.83	**27.22**	23.23
12	AFEW4.0	eNTERFACE	18.68	19.75	17.48	18.10	21.93	21.64	**21.94**	19.93
Avg	-	-	31.51	28.53	26.32	32.86	39.78	40.92	**41.08**	33.87

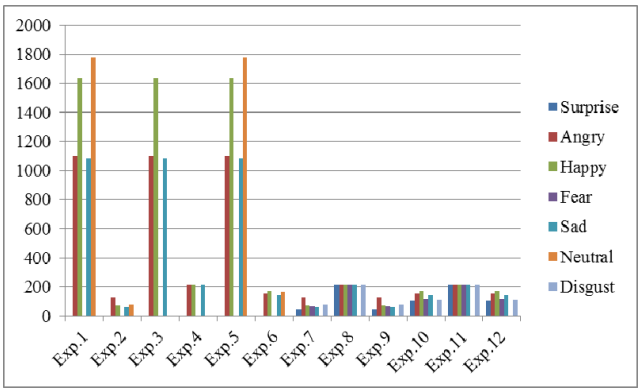

FIGURE 2. Common emotion states used in the twelve emotion states in exp.5 and exp.6 are the same with exp.1 and exp.2. And the common emotion states in exp.7 and exp.8 are angry, disgust, fear, happy and sad. Besides, there are six kind common emotion states in exp.9, exp.10, exp.11 and exp.12, which are angry, disgust, fear, happy, sad and neutral.

coefficient (MFCC), and their 12 functions [33], such as standard deviation and kurtosis, as speech feature representation. As to the evaluation metrics, we employ the weighted average recall (WAR) and the unweighted average recall (UAR) [11] to report the performance of all the methods, which are widely used in cross-corpus speech emotion recognition. WAR is the normal recognition accuracy (i.e., accuracy), while UAR is the mean accuracy of each class (i.e., the accuracy per class divided by the number of classes without considerations of instances per class). Since sample classes are unbalanced in the cross-corpus evaluations, as shown in table 1. It is means that the samples numbers in different classes have a large difference, thus, it is more appropriate to evaluate the results from the perspective of WAR and UAR.

For comparison, we choose KMM [20], KLIEP [9], uLSIF [23], DALSR [6] and DoSL [5] to conduct the experiments under the same protocol as our TRaSL. Besides, we select the linear SVM without any domain adaption ability as the baseline of all the comparison method. The detailed trade-off parameters setting of all the methods in the experiments are listed as follows:

1) For baseline method *SVM*, we use linear kernel function and set C = 1 in the experiments. Meanwhile, for fair comparison, linear kernel function is adopted for all the methods throughout the experiments.
2) For the *KMM* method, there are two important parameters ε and B to be set, which are the upper limit

of importance weight. Depending on the suggestion of [7], the two parameters are set as $B = 1000$ and $\varepsilon = \sqrt{n_{tr}} - \sqrt{1/n_{tr}}$, where n_{tr} denotes the number of training samples.

3) For the *KLIEP* method, no parameter for KLIEP needs to be set.
4) For the *uLSIF*, *DALSR*, *DoSL* and *TRaSL* methods, there are trade-off parameters to be set. In the experiments, grid search strategy is adopted for these cross-corpus speech emotion recognition to build a fair experiment environment, finally the best result is reported with the optimal trade-off parameters. For uLSIF, The trade-off parameter λ search interval is fixed between [1:1:100]$\times$ r (r =1,10,100,100, 1000, 10000, 100000) in the experiments. For DaLSR, it has two important trade-off parameters λ and μ. The optimal values are determined by searching from [1:1:10] for λ and [1:1:10] for μ. For our DoSL and TRaSL, the preset spaces for $\lambda 1$ and $\lambda 2$ are [1:1:10] and [1:1:10], respectively.

Note that the experimental results are directly taken from [5] since the comparative experiments setting are exactly same as that of [5]. Finally, the parameters ($\lambda 1$, $\lambda 2$) of our TRaSL are empirically fixed at (2, 7), (121, 1), (8, 3), (18, 5), (14, 20), and (18, 10) for Exp.7, Exp.8, $\cdots$, Exp.12 experiments, respectively. Meanwhile, we use the method described in Section III-D for TRaSL to predict the emotion labels of target speech samples.

C. RESULTS AND ANALYSIS

In this section, we report the results of the evaluated methods including various DA methods. The experimental results in terms of WAR and UAR of all the methods for all twelve experiments are depicted in Tables 2 and 3, respectively. The normal numbers are the recognition rate and the subscript numbers are the relative rank of UAR and WAR in each method. To observe the influence of different source and target domain on the results and the overall performance of each method, we calculate the average (Avg) results of all the experiments for each method and all the methods in each experiment, which are show in the last row and last column of these tables. From the results, we make the following observations.

Firstly, it can be found that in all experiments, our TRaSL framework achieves promising increases in the performance over the SVM without any domain adaptation ability. Additionally, our TRaSL achieves both best UAR and WAR among all the methods in eight of twelve cases including Exp.5, Exp.7 and Exp.10 to Exp.12. As while, it is clear to see that the UAR of TRaSL in Exp.9 and the WAR of DoSL in Exp.4 are very competitive against the highest results in respective experiments, which is shown in the comparison between KMM and TRaSL in Exp.9 (30.39% v.s. 29.38%) and the comparison between our DoSL and TRaSL in Exp.4 experiment (55.35% v.s. 55.30%).

Secondly, we observe that DALSR outperforms all the comparative methods in terms of WAR and UAR in Exp.8, which shows it is more effective than other methods. Although in the experiments mentioned above, our TRaSL does not perform best in terms of UAR, we can from the results achieved by TRaSL and DALSR (highest), observe that their differences of UAR is actually not large, besides, our TRaSL also achieved highest result of WAR as DALSR. In this case, the UAR and WAR of DALSR are (44.41%, 52.27%), while the results of our TRaSL are (43.44%, 52.71%). In addition, it can be seen that in Exp.3, the results of KMM, KLIEP and uLSIF have big gaps with the baseline method SVM, we guess maybe these three methods have predicted the sample labels to be the same one.

Thirdly, based on our results, it is convincing that the limited label information provided by a small number of samples in source database will lead to low recognition rate. For example, in the cases with eNTERFACE as the target corpus, i.e., Exp.3, Exp.7 and Exp.12, it can be seen that the average performance in terms of WAR and UAR of all the domain adaptation methods can reach 40.84% and 40.93% in Exp.3, in which the source corpus, IEMOCAP, provides large number of samples than the other two experiments. These two metrics drop to 29.77% and 29.76% in Exp.7 and 19.99% and 19.93% in Exp.12, where the source corpus of Exp.7 and Exp.12 are EmoDB and AFEW4.0, respectively. Lastly, a strange phenomenon is found in Exp.1, DALSR and the proposed DoSL and TRaSL achieve WAR of 62.32%, 62.91% and 65.49%, UAR of 62.32%, 62.91% and 61.02%, respectively, which are far higher than the other four comparison methods. Besides, it can be seen at a glance the gap between WAR and UAR of the other four methods is much larger than DALSR, DoSL and TRaSL. Due to the dominant percentage of Anger samples in Emo DB, we consider that most of EmoDB samples may be mistakenly predicted as Angry by the four comparison methods and hence lead to lager gaps between WAR and UAR. Meanwhile, we also find that compared with the experiments of using EmoDB as target database (Exp.1, Exp.8 and Exp.10), there is a big gap between WAR and UAR among all the methods. Besides, one more interesting finding can be obtained according to the tables; the experimental results of using EmoDB as source database are obviously lower than those of using the same database as target one, e.g., Exp3 and Exp.7. It is mostly due to the class imbalance problem existing in EmoDB database, which can be seen in Table 1. In contrast to Exp.1 to Exp.7, and Exp.10, the UAR of our proposed methods in Exp.9 are less than KMM, in which most of the methods achieved low recognition rate. It is probably caused by the unbalance problem of labeled data samples in each class of source corpus.

D. EFFECTIVENESS VERIFICATION

So as to verify the above analysis and further observe how the data imbalance between source and target databases affect the cross-corpus speech emotion recognition tasks, we select three pairs of experiments including Exp.1 to

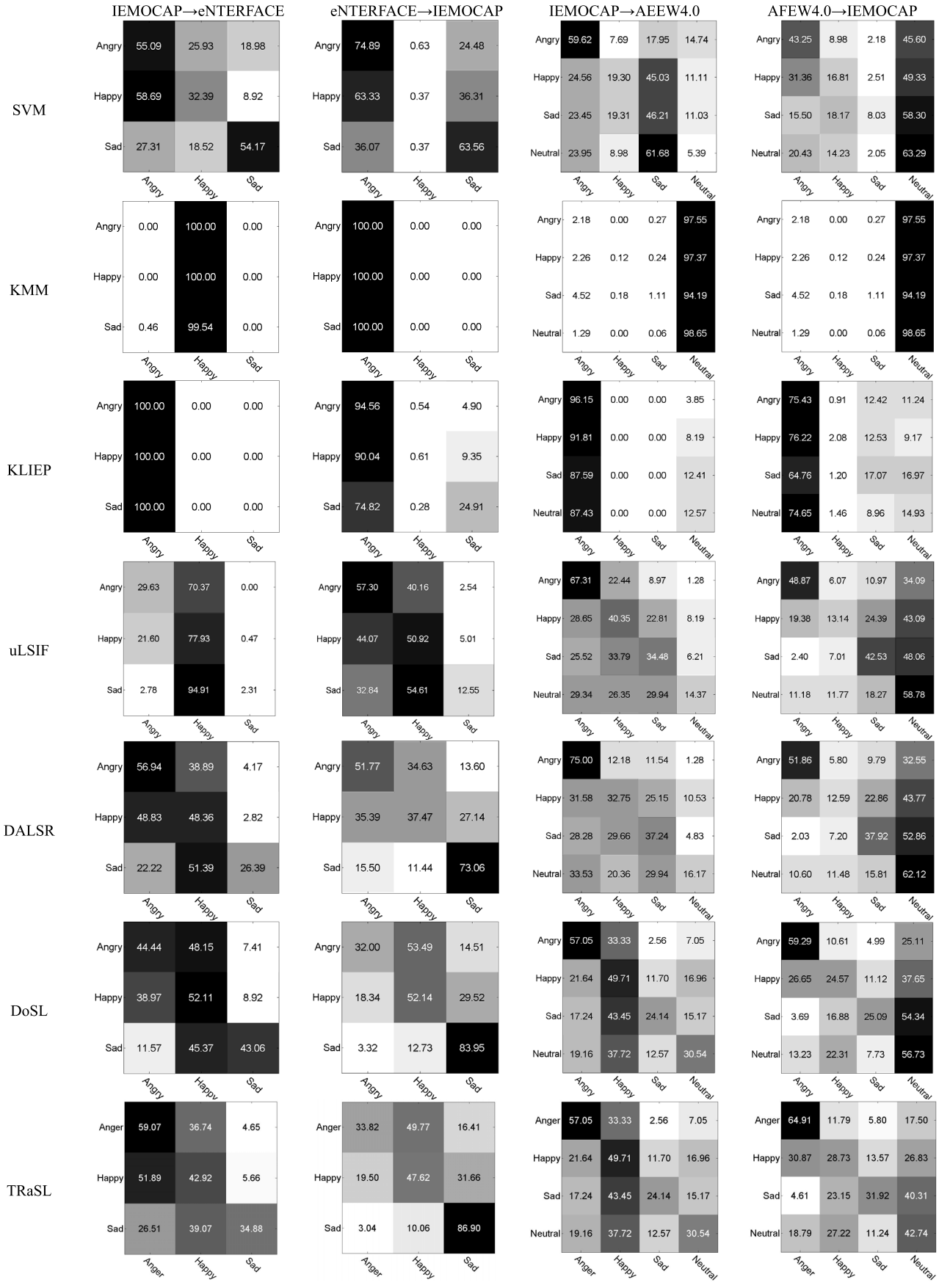

FIGURE 3. The confusion matrices of all the methods in exp.1 to 4, the results are SVM, KMM, KLIEP, uLSIF, DALSR, DoSL and TRaSL, respectively.

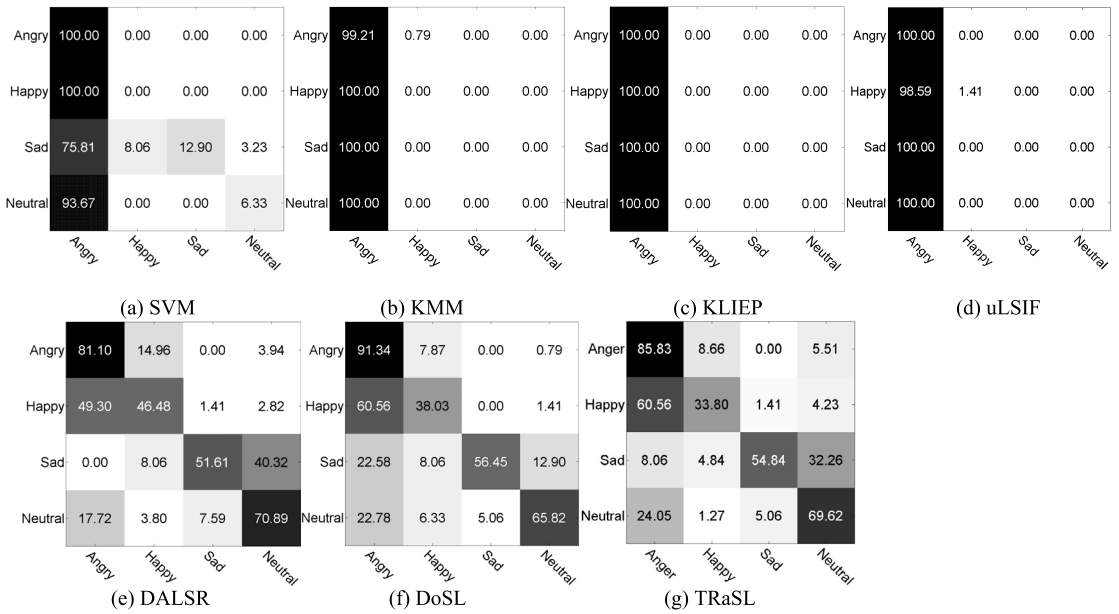

FIGURE 4. The confusion matrices of all the methods in exp.1 from (a) to (g), the results are SVM, KMM, KLIEP, uLSIF, DALSR, DoSL and TRaSL, respectively.

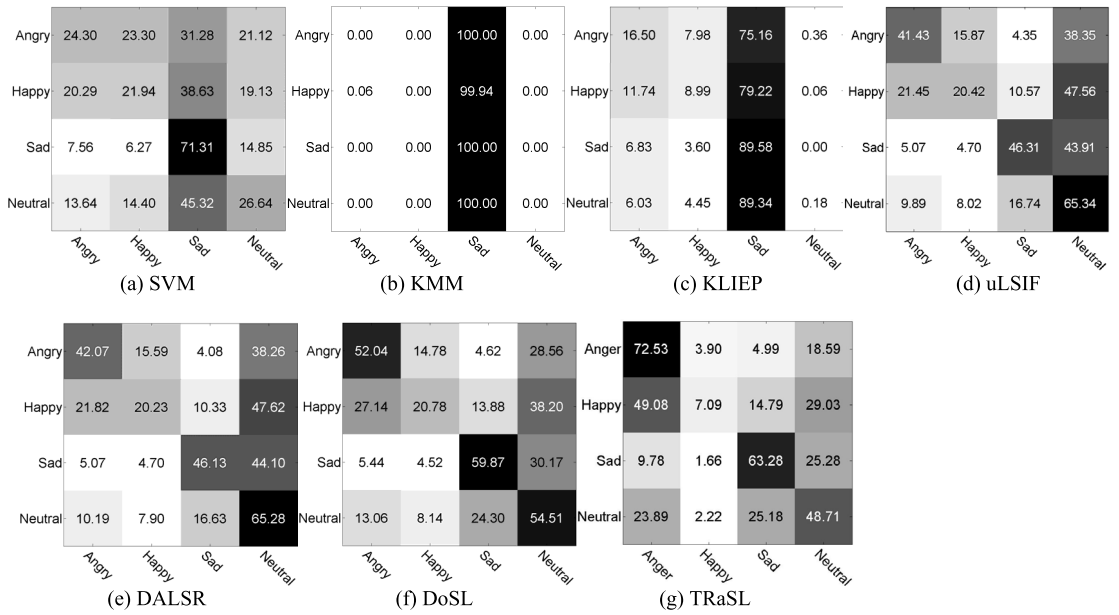

FIGURE 5. The confusion matrices of all the methods in exp.2 from (a) to (g), the results are SVM, KMM, KLIEP, uLSIF, DALSR, DoSL and TRaSL, respectively.

Exp.6, where each database is served as source and target database, respectively, we draw the confusion matrices of all the comparison methods which are depicted in Figs. 3, 4, and 5, respectively. In Fig.3 the left database of "→"is the source database, and the right one is the target database, e.g. IEMOCAP → EmoDB, IEMOCAP is the source database, EmoDB is the target database. From these confusion matrices, some interesting findings can be obtained:

(1) *Performance on Different Emotions:* From the confusion matrix of TRaSL in Figs. 3, 4 and 5, we see that the Angry expression and the Neutral expression are much more easier to be recognized than the other expressions, and the Happy expressions is much more confusing than any other expressions. Additionally, from the confusion matrix of TRaSL in Fig.5, it can be found that there are big gaps between the recognition rate of Anger and Happy expression, where EmoDB and IEMOCAP

are the source and target database, respectively. These experimental results coincide with the analysis from the above experiments.

(2) *The impact of imbalanced database.* By comparing with the confusion matrix of Exp.1, which lies in the first column in Fig. 3, we can clearly see that almost all samples of EmoDB are predicted to be Angry by SVM, KMM, KLIEP and uLSIF, which is confirms our previous analysis. The dominant percentage of Anger samples in EmoDB lead to this phenomenon. Meanwhile, it also explains why there are such big gaps between WAR and UAR in these four methods, which indicates that the proposed TRaSL method is less affected by the extreme class imbalance problem existing in EmoDB and is more applicable to this challenging experiment.

(3) *Performance on limited label information.* Compared the confusion matrix between the last column in Fig. 3 and Fig. 5, it can be found that in Fig. 5 the results of all the methods are seriously affected by the limited label information provided in source database hence most of target samples were wrongly predicted. This is because of that the model cannot get adequate training using small source sample. Though TRaSL method can promisingly alleviate this extremely wrong prediction, we can observe that in Fig. 5 nearly 70% of happy and sad samples are wrongly predicted. Consequently, DA methods including the proposed TRaSL still have very big space for coping with a small number of samples in source database.

TABLE 4. The recognition accuracy in EmoDB to interface.

Methods	Recognition rate (%)					
	Anger	Disgust	Fear	Happy	Sad	WAR
Baseline	*74.42*	*55.39*	*54.13*	*60.03*	*61.02*	*61.42*
TCA[43]	15.75	8.70	26.09	25.35	41.93	22.93
GFK[44]	46.76	26.39	24.07	30.05	24.54	30.36
DR[41]	50.00	37.96	23.61	14.08	33.33	31.85
STM[25]	98.15	0	0	0	23.61	24.42
TNNMF[16]	50.08	29.35	36.92	47.34	46.11	**44.76**
TRaSL	65.74	18.52	26.39	27.70	50.46	37.79

E. FURTHER VERIFICATION

Transfer learning is widely used in many fields. To further illustrate the effectiveness of the proposed TRaSL, and the impact of different features on the algorithms performance. We choose one baseline, which is directly taken from [40], and several state-of-the-art transfer learning methods including DR [41], [42], TCA [43], GFK [44], STM [25],TNNMF [16] to conduct the experiments. In the baseline method the training data and testing data are from the same corpus. In this section, we choose Exp.7 and Exp.8 as the representatives. The WAR results are shown in table 4 and 5. From the tables it clearly to see that, TRaSL achieved better performance than other transfer learning methods.

TABLE 5. The recognition accuracy in interface to EmoDB.

Methods	Recognition rate (%)					
	Anger	Disgust	Fear	Happy	Sad	WAR
Baseline	*73.12*	*81.09*	*68.56*	*53.02*	*79.35*	*71.08*
TCA[43]	33.80	11.11	32.87	15.74	20.83	22.93
GFK[44]	43.31	23.91	17.39	19.72	35.48	30.40
DR[41]	99.21	13.04	1.45	0	29.17	41.07
STM[25]	100.00	0	0	0	0	33.87
TNNMF[16]	35.93	72.07	19.12	24.69	68.97	49.98
TRaSL	81.10	6.52	21.74	12.68	95.16	**52.27**

Besides, we observe that the dimensionality reduction based transfer learning algorithms not achieved excellent effect, i.e., TCA, GFK, DR and STM, which do not take into account the importance of feature selection. Furthermore, the WAR of baseline method is much higher than other methods, which indicate that different feature distributions have great influence on the recognition rate. It also shows the necessity of cross-corpus speech emotion recognition.

F. ABLATION STUDIES

In order to see how the objective function terms affect the performance of TRaSL, ablation studies of the model are investigated. The final objective function is shown in Eq. (4), which is composed of three parts. In this section, we conducted three kinds of experiments.

TRaSL-I: In Eq.(4), $||\mathbf{U}^T||_{2,1}$ term is served as the regularization to select the important features. To proving the impact of it to SER, we removed this term from Eq. (4), and marked it as TRaSL-I.

TRaSL-D: In order to check the effect of discriminant property to the model, the term $||\mathbf{U}^T(\Sigma^s - \Sigma^t)\mathbf{U}||_F^2$, which can keep enough discriminant information is removed, we named it as TRaSL-D.

TRaSL: TRaSL is the final objective function, which includes feature distribution difference term and discriminant property term, meanwhile, importance of feature selection term also takes into account.

TABLE 6. The results of ablation experiments on speech signal feature.

EXP.	Source Corpus	Target Corpus	Method		
			TRaSL-I	TRaSL-D	TRaSL
1	EmoDB	eNTERFACE	27.11	37.51	**37.79**
2	eNTERFACE	EmoDB	22.93	52.00	**52.27**
3	EmoDB	AFEW4.0	21.91	31.00	**31.24**
4	AFEW4.0	EmoDB	28.41	50.00	**50.33**
5	eNTERFACE	AFEW4.0	23.22	24.83	**27.22**
6	AFEW4.0	eNTERFACE	19.03	21.64	**21.94**

We show the ablation experimental on speech signal feature results in Table 6. It can from the results be seen that TRaSL achieved promisingly increase compared with the TRaSL-I and TRaSL-D, which can demonstrate the effectiveness of the proposed TRaSL framework. Furthermore, as the table shows, TRaSL-I gained a lower recognition rate,

which indicated that important feature selection has a great influence on the results.

TABLE 7. The results of ablation experiments on spectrogram features.

EXP.	Source Corpus	Target Corpus	Method		
			TRaSL-I	TRaSL-D	TRaSL
1	EmoDB	IEMOCAP	47.13	47.38	**50.37**
2	IEMOCAP	EmoDB	37.00	57.50	**59.50**
3	EmoDB	IEMOCAP	32.20	41.50	**45.35**
4	IEMOCAP	EmoDB	29.70	27.80	**36.40**
5	EmoDB	IEMOCAP	25.20	35.50	**39.08**
6	IEMOCAP	EmoDB	28.23	47.37	**50.70**

Besides, in order to verify the effectiveness and robustness of the TRaSL, experiments are carried out by using the spectrogram features of emoDB and IEMOCAP databases. For more training data, the utterances are divided into several segments and all segments in the same utterance share the same label. Researchers have point out that a segment longer than 250ms can provide enough emotional information [45], [46]. Similar to [47], in this work, the length of a segment is set to be 265ms. Following the experiment setting in IV-B, we select the samples belonging to the common emotion states from these two datasets, which are served as source and target corpus, alternatively. Meanwhile, we randomly selected three groups of samples from emoDB and IEMOCAP database, respectively. The numbers of samples in each group are (200, 800), (1000, 2000), (3000, 5000). In Table 7, the sample numbers of emoDB and IEMPOCAP in EXP.1 and EXP.2 are 200 and 800, respectively. Similar to EXP.1 and EXP.2, they are 1000, 2000 and 3000, 5000 in EXP.3, EXP.4 and EXP.5, EXP.6. From the results, it can be seen that under the spectrogram-based statistical features our TRaSL also achieved promising results.

G. PARAMETER SENSITIVITY

There are two important trade-off parameters in the proposed TRaSL framework, i.e. λ_1 and λ_2, whose selection will affect the performance of TRaSL. So the next obvious question is, whether the performance of TRaSL is sensitive to the selection of λ_1 and λ_2. To investigate this point, we conduct experiments by fixing the values of one trade-off parameter while changing the other one. As representatives, we select two pairs of experiments including Exp.1, Exp.2, Exp.7 and Exp.8 to conduct the experiments, in which we will report the average recognition accuracy (WAR). The preset spaces are [0.001, 0.01, 0.1, 1, 10, 100] for λ_1 and λ_2. The fixed λ_1 and λ_2 values are consistent with the above experiment in IV-B. The WAR of these parameters are shown in Figs.6. From Figs.6, we can see that the performance of TRaSL varies slightly with respect to the change of λ_1 and λ_2 in all experiments, which indicates that our TRaSL can achieve optimal recognition performance with a wide range of parameter values, i.e., our TRaSL is less sensitive to its trade-off parameters.

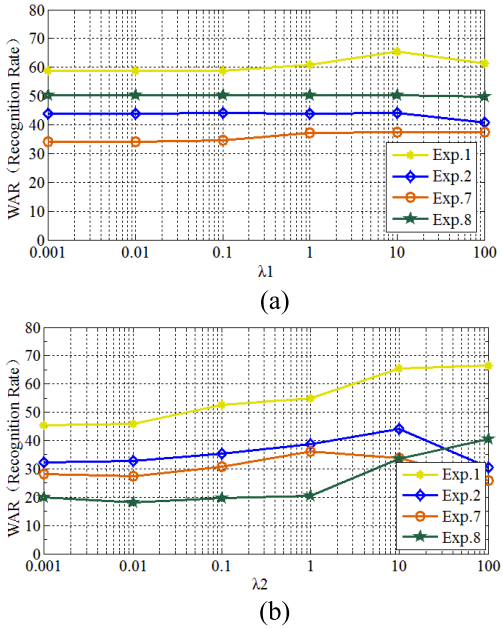

FIGURE 6. Parameter sensitivity experiments for TRaSL. Picture (a) are the average recognition accuracy of TRaSL in exp.1, exp.2, exp.7 and exp.8 with fixed λ_2 when $\lambda_1 \in$ [0.001, 0.01, 0.1, 01, 10, 100], picture (b) are the average recognition accuracy of TRaSL in exp.1, exp.2, exp.7 and exp.8 with fixed λ_1 when $\lambda_2 \in$ [0.001, 0.01, 0.1, 01, 10, 100].

V. CONCLUSION AND FUTURE WORKS

In this work, we propose an unsupervised transfer subspace learning (TRaSL) model via transform the original sample features of source and target database to a predefined common subspace, which can deal with the unsupervised cross-corpus speech emotion recognition (SER) problem. By using TRaSL model, we can learn a projection matrix to transform the source and target speech samples from the original feature space, where the feature distributions of the source and target speech samples have large difference, into the label space, where the transformed source and target speech samples would obey the similar feature distributions. Therefore, the classifier learned based on the transformed labeled source speech samples are then utilized to predict the speech emotion category of the unlabeled target speech samples. Extensive cross-corpus SER experiments based on the four speech emotion corpora are conducted to evaluate the performance of the proposed TRaSL method. The evaluation results demonstrate the superiority of our TRaSL to the recent state-of-the-art cross-corpus SER methods. Besides, the investigations also imply that both the label information provided in source database and the class imbalance of target domain are constraints for domain adaptation, the quantity of label information provided by source database can provide sample emotional features for the model, while the imbalance of target domain will cause the predict result as the same one speech emotion.

In the work, we mainly aim to transform source and target speech signals to share similar feature distributions for FER. It is also expected that a more sophisticated feature selection

method will be designed to further improve the performance. With the development of deep learning techniques, its strong nonlinear representation ability will help bridging the source and target domains. One of our future works will study how to introduce the convolution neutral network into our TRaSL method.

ACKNOWLEDGMENT

The authors would like to thank the comments and suggestions from the editor and reviewers.

REFERENCES

[1] M. E. Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, recognition: Variances and strategies," *IEEE Trans.*, vol. 14, pp. e49–e57, 2006.

[2] Z.-T. Liu, Q. Xie, M. Wu, W.-H. Cao, Y. Mei, and J.-W. Mao, "Speech emotion recognition based on an improved brain emotion learning model," *Neurocomputing*, vol. 309, pp. 145–156, Oct. 2018.

[3] J. Gideon, M. McInnis, and E. M. Provost, "Improving cross-corpus speech emotion recognition with adversarial discriminative domain generalization (ADDoG)," *IEEE Trans. Affect. Comput.*, early access, May 14, 2019, doi: 10.1109/TAFFC.2019.2916092.

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

[5] N. Liu, Y. Zong, B. Zhang, L. Liu, J. Chen, G. Zhao, and J. Zhu, "Unsupervised cross-corpus speech emotion recognition using domain-adaptive subspace learning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5144–5148.

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

[7] Y. Zong, W. Zheng, X. Huang, K. Yan, J. Yan, and T. Zhang, "Emotion recognition in the wild via sparse transductive transfer linear discriminant analysis," *J. Multimodal User Interface*, vol. 10, no. 2, pp. 163–172, Jun. 2016.

[8] C. C. Chang and C. J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, 2011.

[9] M. Sugiyama, S. Nakajima, H. Kashima, P. V. Buenau, and M. Kawanabe, "Direct importance estimation with model selection and its application to covariate shift adaptation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2008, pp. 1433–1440.

[10] S. Mariooryad and C. Busso, "Exploring cross-modality affective reactions for audiovisual emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 4, no. 2, pp. 183–196, Apr. 2013.

[11] B. Schuller, B. Vlasenko, F. Eyben, M. Wollmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll, "Cross-corpus acoustic emotion recognition: Variances and strategies," *IEEE Trans. Affect. Comput.*, vol. 1, no. 2, pp. 119–131, Jul./Dec. 2010.

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

[13] K. Yan, W. Zheng, Z. Cui, and Y. Zong, "Cross-database facial expression recognition via unsupervised domain adaptive dictionary learning," in *Proc. Int. Conf. Neural Inf. Process.*, 2016, pp. 427–434.

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

[15] A. Hassan, R. Damper, and M. Niranjan, "On acoustic emotion recognition: Compensating for covariate shift," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 7, pp. 1458–1468, Jul. 2013.

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

[17] Z. Huang, W. Xue, Q. Mao, and Y. Zhan, "Unsupervised domain adaptation for speech emotion recognition using PCANet," *Multimedia Tools Appl.*, vol. 76, pp. 6785–6799, Feb. 2016.

[18] H. Sagha, J. Deng, M. Gavryukova, J. Han, and B. Schuller, "Cross lingual speech emotion recognition using canonical correlation analysis on principal component subspace," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 5800–5804.

[19] N. Liu, B. Zhang, Y. Zong, L. Liu, J. Chen, G. Zhao, and L. Zhu, "Super wide regression network for unsupervised cross-database facial expression recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 1897–1901.

[20] S. Jialin Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[21] J. Huang, A. Gretton, K. M. Borgwardt, B. Schölkopf, and A. J. Smola, "Correcting sample selection bias by unlabeled data," in *Proc. Adv. Neural Inf. Process. Syst.*, 2006, pp. 601–608.

[22] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 915–928, Jun. 2007.

[23] T. Kanamori, S. Hido, and M. Sugiyama, "A least-squares approach to direct importance estimation," *J. Mach. Learn. Res.*, vol. 10, pp. 1391–1445, Jul. 2009.

[24] W.-S. Chu, F. De la Torre, and J. F. Cohn, "Selective transfer machine for personalized facial action unit detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3515–3522.

[25] W.-S. Chu, F. De la Torre, and J. F. Cohn, "Selective transfer machine for personalized facial expression analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 3, pp. 529–545, Mar. 2017.

[26] K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Scholkopf, and A. J. Smola, "Integrating structured biological data by kernel maximum mean discrepancy," *Bioinformatics*, vol. 22, no. 14, pp. e49–e57, Jul. 2006.

[27] W. Zheng, M. Xin, X. Wang, and B. Wang, "A novel speech emotion recognition method via incomplete sparse least square regression," *IEEE Signal Process. Lett.*, vol. 21, no. 5, pp. 569–572, May 2014.

[28] A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola, "A kernel method for the two-sample-problem," in *Proc. Adv. Neural Inf. Process. Syst.*, 2006, pp. 513–520.

[29] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 171–184, Jan. 2013.

[30] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. Interspeech*, vol. 5, 2005, pp. 1517–1520.

[31] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACE'05 audio-visual emotion database," in *Proc. 22nd Int. Conf. Data Eng. Workshops (ICDEW)*, Apr. 2006, p. 8.

[32] A. Dhall, R. Goecke, J. Joshi, K. Sikka, and T. Gedeon, "Emotion recognition in the wild challenge 2014: Baseline, data and protocol," in *Proc. 16th Int. Conf. Multimodal Interact.*, 2014, pp. 461–466.

[33] B. W. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in *Proc. Interspeech*, 2009, pp. 312–315.

[34] C. Busso, M. Bulut, C. C. Lee, and A. Kazemzadeh, "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, p. 335, 2008.

[35] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: The munich versatile and fast open-source audio feature extractor," in *Proc. 18th ACM Int. Conf. Multimedia*, 2010, pp. 1459–1462.

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

[37] Y. Kim, H. Lee, and E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 3687–3691.

[38] M. Shah, C. Chakrabarti, and A. Spanias, "A multi-modal approach to emotion recognition using undirected topic models," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, Jun. 2014, pp. 754–757.

[39] Y. Zong, W. Zheng, X. Huang, J. Shi, Z. Cui, and G. Zhao, "Domain regeneration for cross-database micro-expression recognition," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2484–2498, May 2018.

[40] P. Song and W. Zheng, "Feature selection based transfer subspace learning for speech emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 11, no. 3, pp. 373–382, Sep. 2020.

[41] S. J. Pan, J. T. Kwok, and Q. Yang, "Transfer learning via dimensionality reduction," in *Proc. AAAI Conf. Artif. Intell.*, 2008, pp. 677–682.

[42] P. Song, Y. Jin, M. Xin, and L. Zhao, "Speech emotion recognition using transfer learning," *IEICE Trans. Inf. Syst.*, vol. 97, no. 9, pp. 2530–2532, Sep. 2014.

[43] S. Jialin Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Trans. Neural Netw.*, vol. 22, no. 2, pp. 199–210, Feb. 2011.
[44] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 2066–2073.
[45] Y. Kim and E. M. Provost, "Emotion classification via utterance-level dynamics: A pattern-based approach to characterizing affective expressions," in *Proc. ICASSP*, May 2013, pp. 3677–3681.
[46] E. M. Provost, "Identifying salient sub-utterance emotion dynamics using flexible units and estimates of affective flow," in *Proc. ICASSP*, May 2013, pp. 3682–3686.
[47] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Proc. INTERSPEECH*, Jul. 2014, pp. 223–227.

NA LIU received the B.S. degree in automation from Tianjin University, Tianjin, China, in 2011, and the M.S. and Ph.D. degrees in detection technology and automatic equipment and computer science and technology from the Tianjin University of Technology, Tianjin, in 2014 and 2019, respectively. She was a Visiting Student with the Center of Machine Vision and Signal Analysis, University of Oulu, Finland, from March 2017 to February 2018. She is currently a Lecturer with the Tianjin Key Laboratory for Control Theory and Applications in Complicated System, School of Electrical and Electronic Engineering, Tianjin University of Technology. Her research interests include affective computing, computer vision, and artificial intelligence.

BAOFENG ZHANG received the Ph.D. degree in measuring and testing technologies and instruments from Tianjin University, Tianjin, China, in 2002. He was working as a Visiting Scholar with Physikalisch-Technische Bundesanstalt (PTB), Germany, from December 1988 to January 2001. He is specially appointed as a Professor of the Higher Learning Institutions, Tianjin. He is currently a Professor with the Tianjin Key Laboratory for Control Theory and Applications in Complicated System, School of Electrical and Electronic Engineering, Tianjin University of Technology. His research interests include vision detection, motion control systems, pattern recognition, and computer vision.

BIN LIU received the Ph.D. degree in instrument science from Tianjin University, Tianjin, in 2010. He was working as a Visiting Scholar with University College Dublin, Ireland, from September 2018 to August 2019. He is currently an Associate Professor with the School of Electrical and Electronic Engineering, Tianjin University of Technology. His research interests include computer vision and photoelectric detection technology.

JINGANG SHI (Member, IEEE) received the B.S. and Ph.D. degrees from the Department of Electronics and Information Engineering, Xi'an Jiaotong University, China. From 2017 to 2020, he was a Postdoctoral Researcher with the Center for Machine Vision and Signal Analysis, University of Oulu, Finland. Since 2020, he has been an Associate Professor with the School of Software, Xi'an Jiaotong University. His current research interests include image restoration, face analysis, and biomedical signal processing.

LEI YANG received the M.S. degree in detection technology and automatic equipment from the Tianjin University of Technology, Tianjin, China, in 2018. He is currently a System Engineer with Autobrain (Tianjin) Technology Company Ltd. His research interests include automatic driving, computer vision, vision detection, and measuring and testing technologies.

ZHIWEI LI received the Ph.D. degree in instrument science and technology from Tianjin University, Tianjin, China, in 2016. He is currently a Lecturer with the Tianjin Key Laboratory for Control Theory and Applications in Complicated System, School of Electrical and Engineering, Tianjin University of Technology, Tianjin. His research interests include vision detection and measuring and testing technologies. He is a member of the Chinese Society for Measurement.

JUNCHAO ZHU received the Ph.D. degree in measuring and testing technologies and instruments from Tianjin University, Tianjin, China, in 2007. He is currently a Professor with the Tianjin Key Laboratory for Control Theory and Applications in Complicated System, School of Electrical and Electronic Engineering, Tianjin University of Technology. His research interests include photoelectric detection technology, embedded measurement and control systems, vision detection, and computer vision.