Seeking Salient Facial Regions for Cross-Database Micro-Expression Recognition

Xingxun Jiang, Yuan Zong[#], Wenming Zheng[#], Jiateng Liu, Mengting Wei
Key Laboratory of Child Development and Learning Science of Ministry of Education,
School of Biological Science and Medical Engineering, Southeast University, Nanjing 210096, China
{jiangxingxun, xhzongyuan, wenming_zheng, jiateng_liu, weimengting}@seu.edu.cn

***Abstract*—Cross-Database Micro-Expression Recognition (CDMER) aims to develop the Micro-Expression Recognition (MER) methods with strong domain adaptability, i.e., the ability to recognize the Micro-Expressions (MEs) of different subjects captured by different imaging devices in different scenes. The development of CDMER is faced with two key problems: 1) the severe feature distribution gap between the source and target databases; 2) the feature representation bottleneck of ME such local and subtle facial expressions. To solve these problems, this paper proposes a novel Transfer Group Sparse Regression method, namely TGSR, which selects those salient facial regions for CDMER and performs sparse operations at the feature level in order to 1) optimize the difference measurement between the source and target databases so as to better alleviate this difference, 2) highlight the valid part to make extracted feature more effective and explicable. We use two public micro-expression databases, i.e., CASME II and SMIC, to evaluate the proposed TGSR method. Experimental results show that the proposed TGSR learns explicable and highly discriminative ME features, and outperforms most state-of-the-art subspace-learning-based domain-adaptive methods for CDMER.**

I. Introduction

Micro-Expression (ME) is a low amplitude and short duration facial expression which may reflect subjects' genuine emotions[1], [2].It is indispensable in many fields, such as criminal investigations[3], clinical diagnosis[4], human-computer interaction[5], [6], [7], [8], etc.

Due to the huge potential value of MEs, many efforts have been made to design an automatic Micro-Expression Recognition (MER) system over last few decades[9], [10], [11], [12], [13], [14], [15], [16], [17]. In traditional pattern recognition methods, feature extraction is the most important step. Researchers focused on this and proposed many well-designed hand-crafted features, such as LBP-TOP[18], LBP-SIP[10], FDM[15], etc. These proposed features are both beneficial to improving MER performance. In deep learning methods[19], [20], [21], [22], researchers integrate the feature extraction and classification into an end-to-end approach. It learns more effective features and obtains better performance through the deep and hierarchical learning approach. Past research has promoted the rapid development of automated MER technology. However, most existing methods are only evaluated in one single database, which may make the performance of recognizing micro-expressions from different domains, i.e., the MEs captured by different imaging devices and different subjects in different scenes, out of satisfaction.

In order to learn a domain-robust MER model, researchers have turned their interests to domain-adaptive MER methods in recent years. A new challenging topic has thus emerged, i.e., Cross-Database Micro-Expression Recognition (CDMER). It mimics the domain variation problem in practical application and evaluates the method's adaptive ability by the operation of training the model in one micro-expression database (source database) and testing in the other one (target database). CDMER[23], [24], [25] is faced with two problems: 1) the severe feature distribution gap between the source and target databases, and 2) the feature representation bottleneck of ME such a subtle and local facial expressions.

It has been widely validated that salient region selection benefits face-related visual tasks. Inspired by this, this paper selects salient regions from the whole face in order to 1) optimize the difference measurement between the source and target databases so as to alleviate this difference better; 2) highlight the valid part to make extracted feature more discriminative and explicable. We propose a novel Transfer Group Sparse Regression method (TGSR) which introduces a learnable binary sparse regression matrix shared between the source and target databases to implement the salient facial region selection for CDMER. Specially, TGSR contains three terms: a regression term for bridging micro-expression features and labels, a Frobenius norm sparse term for sparsing the learnable regression matrix, and a joint feature distribution term for measuring the difference between the source and target databases. We hope the proposed TGSR model can learn a more discriminative CDMER model by jointly optimizing these three terms. We evaluate our method on CASME II[26] and SMIC[27] databases. Experimental results and corresponding visualization demonstrate that our proposed TGSR can effectively solve these two problems and outperforms most state-of-the-art subspace-learning-based domain-adaptive methods for CDMER.

II. Method

A. The Generation of Micro-Expression Features

Extracting ME feature is the first step of CDMER. Firstly, we use the grid-based multi-scale spatial division scheme[28] as Fig. 1 shown to divide the cropped ME sequence into K spatial local sequences. Then we extracted d-dimensional

# indicates the corresponding authors

hand-crafted spatio-temporal feature $\boldsymbol{x}_k$ of each spatial local sequence and obtain the hierarchical feature $\boldsymbol{x}^\nu = \left[\boldsymbol{x}_1^{\mathrm{T}}, \cdots, \boldsymbol{x}_K^{\mathrm{T}}\right]^{\mathrm{T}} \in \mathbb{R}^{Kd}$ of the ME sequence by concatenating these spatial local features one by one. Suppose that we have N_s source and N_t target micro-expression samples, the feature matrix of source and target database can be denoted as $\boldsymbol{X}^s = \left[\boldsymbol{X}_1^{s\mathrm{T}}, \cdots, \boldsymbol{X}_K^{s\mathrm{T}}\right]^{\mathrm{T}} \in \mathbb{R}^{Kd\times N_s}$ and $\boldsymbol{X}^t = \left[\boldsymbol{X}_1^{t\mathrm{T}}, \cdots, \boldsymbol{X}_K^{t\mathrm{T}}\right]^{\mathrm{T}} \in \mathbb{R}^{Kd\times N_t}$. Here, each column of $\boldsymbol{X}^s$ and $\boldsymbol{X}^t$ is a feature vector like $\boldsymbol{x}^\nu$, they respectively denote the micro-expression feature from the source and target databases. $\boldsymbol{X}_i^s \in \mathbb{R}^{d\times N_s}$ and $\boldsymbol{X}_i^t \in \mathbb{R}^{d\times N_t}$ respectively denote the feature matrix of i-th spatial local sequence from the source and target databases. The label of source micro-expression database is denoted by $\boldsymbol{L}^s = [\boldsymbol{l}_1^s, \cdots, \boldsymbol{l}_{N_s}^s] \in \mathbb{R}^{C\times N_s}$, where C is the total category number and the j-th column of $\boldsymbol{L}^s$ denotes the label vector of j-th source micro-expression sample. The label vector of j-th sample $\boldsymbol{l}_j^s = [l_{j,1}^s, \cdots, l_{j,C}^s]^{\mathrm{T}}$ is a one-hot vector in which only one element $l_{j,c}^s$ equals one and the others are zero. It indicates that j-th sample from the source database belongs to c-th micro-expression category.

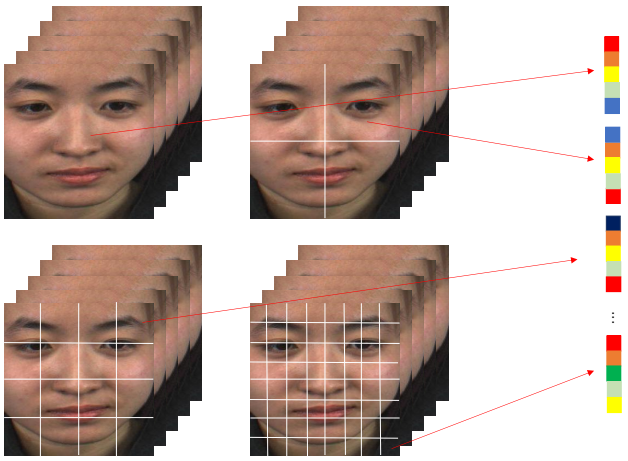

Fig. 1. The grid-based multi-scale spatial division scheme for extracting micro-expression features.

B. Proposed Method

The basic idea of our proposed TGSR is learning a linear regression matrix $\boldsymbol{C} = [\boldsymbol{C}_1^{\mathrm{T}}, ..., \boldsymbol{C}_K^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{Kd\times C}$ shared between the source and target databases to construct the relation between micro-expression features and labels, the relation between source and target databases, and select those salient facial regions, which can be formulated as Equ. (1),

$$\min_{\boldsymbol{C}_i} \left\| \boldsymbol{L}^s - \sum_{i=1}^{K} \boldsymbol{C}_i^{\mathrm{T}} \boldsymbol{X}_i^s \right\|_F^2 + \xi f_1(\boldsymbol{C}_i) + \lambda f_2(\boldsymbol{C}_i), \quad (1)$$

where $\boldsymbol{C}_i \in \mathbb{R}^{C\times d}$ is such a domain-invariant regression matrix to bridge the features of i-th facial region with the corresponding sample labels, $f_1(\boldsymbol{C}_i)$ and $f_2(\boldsymbol{C}_i)$ are the well-designed regularization, and ξ and λ are corresponding weighting hyper-parameters.

Regularization $f_1(\boldsymbol{C}_i)$ is called the database difference elimination one, which measures the difference between the source and target micro-expression databases. By minimizing $f_1(\boldsymbol{C}_i)$ together with the regression term, we can alleviate this difference. We use the maximum mean discrepancy (MMD) to serve as this regularization term, which can be expressed as Equ. (2),

$$MMD\left(\boldsymbol{X}^s, \boldsymbol{X}^t\right) = \left\| \frac{1}{N_s} \sum_{i=1}^{K} \Phi\left(\boldsymbol{X}_i^s\right) \mathbf{1}_s - \frac{1}{N_t} \sum_{i=1}^{K} \Phi\left(\boldsymbol{X}^t\right) \mathbf{1}_t \right\|_{\mathcal{H}}, \quad (2)$$

where $\Phi(\cdot)$ is a kernel mapping operator projecting micro-expression features from the original space to an infinite one, $\mathbf{1}_s \in \mathbb{R}^{N_s}$ and $\mathbf{1}_t \in \mathbb{R}^{N_t}$ are the vectors filled with scalar value one which used to convert the source and target feature into scalar values respectively. However, the kernel mapping operator is unsolvable, so we further modify the MMD into Equ. (3) to serve as $f_1(\boldsymbol{C}_i)$,

$$f_1\left(\boldsymbol{C}_i\right) = \left\| \frac{1}{N_s} \sum_{i=1}^{K} \boldsymbol{C}_i^{\mathrm{T}} \boldsymbol{X}_i^s \mathbf{1}_s - \frac{1}{N_t} \sum_{i=1}^{K} \boldsymbol{C}_i^{\mathrm{T}} \boldsymbol{X}_i^t \mathbf{1}_t \right\|_F^2. \quad (3)$$

By relaxing the difference measurement involving kernel-mapped feature into the label space difference, Equ. (2) becomes solvable. Meanwhile, we optimize this measurement by improving extracted ME features using salient facial region selection so as to alleviate this difference better. Regularization $f_2(\boldsymbol{C}_i)$ is a sparse term designed to select the salient facial regions in order to promote the extracted ME feature more effective and explicable, which is defined as Equ. (4),

$$f_2\left(\boldsymbol{C}_i\right) = \lambda \sum_{i=1}^{K} \left\|\boldsymbol{C}_i\right\|_F. \quad (4)$$

By substituting Equ. (3) and Equ. (4) into Equ. (1), we can rewrite the objective function into Equ. (5),

$$\begin{aligned} \min_{\boldsymbol{C}_i} & \left\| \boldsymbol{L}^s - \sum_{i=1}^{K} \boldsymbol{C}_i^{\mathrm{T}} \boldsymbol{X}_i^s \right\|_F^2 + \lambda \sum_{i=1}^{K} \left\|\boldsymbol{C}_i\right\|_F \\ & + \xi \left\| \frac{1}{N_s} \sum_{i=1}^{K} \boldsymbol{C}_i^{\mathrm{T}} \boldsymbol{X}_i^s \mathbf{1}_s - \frac{1}{N_t} \sum_{i=1}^{K} \boldsymbol{C}_i^{\mathrm{T}} \boldsymbol{X}_i^t \mathbf{1}_t \right\|_F^2. \end{aligned} \quad (5)$$

C. Optimization

We can use the Alternative Direction Method (ADM)[29] and Inexact Augmented Lagrangian Multiplier (IALM)[30] to solve Equ. (5). We firstly rewrite Equ. (5) into Equ. (6),

$$\min_{\boldsymbol{C}_i} \left\| \tilde{\boldsymbol{L}} - \sum_{i=1}^{K} \boldsymbol{C}_i^{\mathrm{T}} \tilde{\boldsymbol{X}}_i \right\|_F^2 + \lambda \sum_{i=1}^{K} \left\|\boldsymbol{C}_i\right\|_F, \quad (6)$$

where $\tilde{\boldsymbol{L}} = [\boldsymbol{L}^s, \mathbf{0}]$, $\mathbf{0} \in \mathbb{R}^{C\times 1}$, $\tilde{\boldsymbol{X}}_i = \left[\boldsymbol{X}_i^s, \sqrt{\xi}(\frac{1}{N_s}\boldsymbol{X}_i^s\mathbf{1}_s - \frac{1}{N_t}\boldsymbol{X}_i^t\mathbf{1}_t)\right]$. Then we introduce a

new variable $\boldsymbol{D} = [\boldsymbol{D}_1^{\mathrm{T}}, \cdots, \boldsymbol{D}_K^{\mathrm{T}}]^{\mathrm{T}}$ equals to variable $\boldsymbol{C} = [\boldsymbol{C}_i^{\mathrm{T}}, \cdots, \boldsymbol{C}_K^{\mathrm{T}}]^{\mathrm{T}}$, and convert the optimization of Equ. (6) into a constrained one as Equ. (7),

$$\begin{aligned} &\min_{\boldsymbol{C},\boldsymbol{D}} \left\| \tilde{\boldsymbol{L}} - \sum_{i=1}^{K} \boldsymbol{D}_i^{\mathrm{T}} \tilde{\boldsymbol{X}}_i \right\|_F^2 + \lambda \sum_{i=1}^{K} \|\boldsymbol{C}_i\|_F, \\ &\text{s.t. } \boldsymbol{D}_i = \boldsymbol{C}_i. \end{aligned} \tag{7}$$

Subsequently, we can obtain the corresponding augmented Lagrange function as Equ. (8) shown,

$$\begin{aligned} \Gamma(\boldsymbol{C}_i, \boldsymbol{D}_i, \boldsymbol{P}_i, \mu) = &\left\| \tilde{\boldsymbol{L}} - \sum_{i=1}^{K} \boldsymbol{D}_i^{\mathrm{T}} \tilde{\boldsymbol{X}}_i \right\|_F^2 + \lambda \sum_{i=1}^{K} \|\boldsymbol{C}_i\|_F \\ &+ \sum_{i=1}^{K} \mathrm{tr}\left[\boldsymbol{P}_i^{\mathrm{T}} (\boldsymbol{C}_i - \boldsymbol{D}_i)\right] + \frac{\mu}{2} \sum_{i=1}^{K} \|\boldsymbol{C}_i - \boldsymbol{D}_i\|_F^2, \end{aligned} \tag{8}$$

where $\boldsymbol{P}_i \in \mathbb{R}^{d \times C}$ denotes the Lagrangian multiplier matrix corresponding to the i-th facial spatial local sequence, and μ is the weighting hyper-parameter.

We can obtain the optimal solution $\hat{\boldsymbol{C}}_i$ of $\boldsymbol{C}_i$ when minimizing the Lagrange function of Equ. (8) by iteratively update $\boldsymbol{C}_i$ and $\boldsymbol{D}_i$. Specifically, we need to repeat the following four steps:

1) Fix $\boldsymbol{C}$, $\boldsymbol{P}$, μ and update $\boldsymbol{D}$:

In this step, the optimization problem with respect to the sub-matrix $\boldsymbol{D}_i$ of $\boldsymbol{D}$ can be written as Equ. (9),

$$\min_{\boldsymbol{D}} \left\| \tilde{\boldsymbol{L}} - \boldsymbol{D}^{\mathrm{T}} \tilde{\boldsymbol{X}} \right\|_F^2 + \mathrm{tr}\left[\boldsymbol{P}^{\mathrm{T}} (\boldsymbol{C} - \boldsymbol{D})\right] + \frac{\mu}{2} \|\boldsymbol{C} - \boldsymbol{D}\|_F^2, \tag{9}$$

where $\boldsymbol{P}^{\mathrm{T}} = [\boldsymbol{P}_1^{\mathrm{T}}, \cdots, \boldsymbol{P}_K^{\mathrm{T}}]$, $\boldsymbol{P} \in \mathbb{R}^{Kd \times C}$, $\boldsymbol{P}_j \in \mathbb{R}^{d \times C}$. The closed-form solution of Equ. (9) as Equ. (12) shows.

2) Fix $\boldsymbol{D}$, $\boldsymbol{P}$, μ update $\boldsymbol{C}$:

In this step, the optimization problem with respect to the sub-matrix $\boldsymbol{C}_i$ of $\boldsymbol{C}$ can be written as Equ. (10),

$$\begin{aligned} \min_{\boldsymbol{C}_i} \lambda \sum_{i=1}^{K} \|\boldsymbol{C}_i\|_F + &\sum_{i=1}^{K} \mathrm{tr}\left[\boldsymbol{P}_i^{\mathrm{T}} (\boldsymbol{C}_i - \boldsymbol{D}_i)\right] \\ &+ \frac{\mu}{2} \sum_{i=1}^{K} \|\boldsymbol{C}_i - \boldsymbol{D}_i\|_F^2. \end{aligned} \tag{10}$$

We can convert Equ. (10) into Equ. (11), and obtain the optimal $\boldsymbol{C}$ using Equ. (13).

$$\min_{\boldsymbol{C}_i} \sum_{i=1}^{K} \left(\frac{\lambda}{\mu} \|\boldsymbol{C}_i\|_F + \frac{1}{2} \left\| \boldsymbol{C}_i - (\boldsymbol{D}_i - \frac{\boldsymbol{P}_i}{\mu}) \right\|_F^2 \right) \tag{11}$$

3) Update $\boldsymbol{P}$ and μ.

4) Check the convergence of $\|\boldsymbol{C} - \boldsymbol{D}\|_\infty < \varepsilon$.

Algorithm 1 shows more optimization details.

Algorithm 1 The Algorithm for the optimal solution of regression matrix $\boldsymbol{C}$ in TGSR method.

Input: Data matrix $\tilde{\boldsymbol{L}}$ and $\tilde{\boldsymbol{X}} = [\tilde{\boldsymbol{X}}_1^{\mathrm{T}}, \cdots, \tilde{\boldsymbol{X}}_K^{\mathrm{T}}]^{\mathrm{T}}$, the salient facial region number κ, the scalar parameter ρ, μ_{max}.

- Initializing the regression matrix $\boldsymbol{C} = [\boldsymbol{C}_1^{\mathrm{T}}, \cdots, \boldsymbol{C}_K^{\mathrm{T}}]^{\mathrm{T}}$
- Initializing the Lagrangian multiplier matrix $\boldsymbol{P} = [\boldsymbol{P}_1^{\mathrm{T}}, \cdots, \boldsymbol{P}_K^{\mathrm{T}}]^{\mathrm{T}}$ and the weighting coefficient μ.

Repeating steps 1) to 4) until convergence.

1: Fix $\boldsymbol{C}, \boldsymbol{P}, \mu$ and update $\boldsymbol{D}$:

$$\boldsymbol{D} = \left(\mu \boldsymbol{I}_{Kd} + 2\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^{\mathrm{T}}\right)^{-1} \left(2\tilde{\boldsymbol{X}}\tilde{\boldsymbol{L}}^{\mathrm{T}} + \boldsymbol{P} + \mu \boldsymbol{C}\right), \tag{12}$$

where $\boldsymbol{I}_{Kd} \in \mathbb{R}^{Kd \times Kd}$ is the identity matrix.

2: Fix $\boldsymbol{D}, \boldsymbol{P}, \mu$ and update $\boldsymbol{C}$:

Calculate $d_i = \left\| \boldsymbol{D}_i - \frac{\boldsymbol{P}_i}{\mu} \right\|_F$, and sort the value of d_i, such that $d_{i_1} \geq d_{i_2} \geq \cdots \geq d_{i_K}$, Let $\lambda = \mu d_{i_{\kappa+1}}$, then update $\boldsymbol{C}$ according to

$$\boldsymbol{C}_i = \begin{cases} \frac{d_i - \frac{\lambda}{\mu}}{d_i} (\boldsymbol{D}_i - \frac{\boldsymbol{P}_i}{\mu}), & \frac{\lambda}{\mu} < d_i, \\ \boldsymbol{0}, & \frac{\lambda}{\mu} \geq d_i. \end{cases} \tag{13}$$

3: Update $\boldsymbol{P}$ and μ:

$$\boldsymbol{P} = \boldsymbol{P} + \mu(\boldsymbol{D} - \boldsymbol{C}), \ \mu = \min(\rho\mu, \mu_{max})$$

4: Check convergence:

$$\|\boldsymbol{C} - \boldsymbol{D}\|_\infty < \varepsilon$$

Output: The solution $\hat{\boldsymbol{C}}$ of regression matrix $\boldsymbol{C}$.

D. Application for CDMER

Based on the labeled source and the unlabeled target databases, we can solve the optimal solution $\hat{\boldsymbol{C}}$ of regression matrix $\boldsymbol{C}$ using aforementioned optimization method. Then, we can extract the feature $\boldsymbol{x}_i^{te} \in \mathbb{R}^{Kd}$ of micro-expression samples to be predicted and predict the label vector $\boldsymbol{l}^{te}$ by solving the optimization problem as Equ. (14),

$$\begin{aligned} &\min_{\boldsymbol{l}^{te}} \left\| \boldsymbol{l}^{te} - \sum_{i=1}^{K} \hat{\boldsymbol{C}}_i^{\mathrm{T}} \boldsymbol{x}_i^{te} \right\|_F^2, \\ &\text{s.t. } \quad \boldsymbol{l}^{te} \geq 0.\boldsymbol{1}^{\mathrm{T}} \boldsymbol{l}^{te} = 1, \end{aligned} \tag{14}$$

where $\hat{\boldsymbol{C}}_i \in \mathbb{R}^{d \times C}$ is the optimal solution of the regression matrix for the i-th facial spatial local region, and $\hat{\boldsymbol{C}}^{\mathrm{T}} = \left[\hat{\boldsymbol{C}}_1^{\mathrm{T}}, \cdots, \hat{\boldsymbol{C}}_K^{\mathrm{T}}\right]$, $\hat{\boldsymbol{C}}^{\mathrm{T}} \in \mathbb{R}^{C \times Kd}$, $\boldsymbol{l}^{te} \in \mathbb{R}^{C}$. Then we can use $\hat{c} = \arg\max_j \left\{\boldsymbol{l}_j^{te}\right\}$ to assign it to the largest entry index of the predicted label vector, i.e., micro-expression category $\hat{c}$.

III. Experiment

A. Experiment Setup

Database. We evaluated our method on Selected CASME II and SMIC database. CASME II[26] contains 255 micro-expression samples from 26 subjects with seven category micro-expressions, i.e., *Disgust*, *Fear*, *Happiness*, *Others*, *Repression*, *Sadness*, and *Surprise*. We selected the samples

TABLE I
THE STATISTICS OF SELECTED CASME II AND SMIC DATABASE.

Dataset	Category		
	Positive	Negative	Surprise
Selected CASME II	32	73	25
SMIC-HS	51	70	43
SMIC-VIS	23	28	20
SMIC-NIR	23	28	20

of *Disgust*, *Happiness*, *Repression*, and *Surprise* to be the Selected CASME II. SMIC[27] records 306 micro-expression samples from 16 subjects in three modalities with three category micro-expressions, i.e., *Positive*, *Negative*, and *Surprise*. The SMIC-HS subset contains 164 micro-expression samples captured by a high-speed camera at 100 frames/s. The SMIC-VIS subset contains 71 micro-expression samples captured by a general visual camera at 25 frames/s. The SMIC-NIR subset contains 71 micro-expression samples captured by a near-infrared camera. In order to make the Selected CASME II and SMIC databases share the same label categories, we converted the labels of Selected CASME II: relabelled the label *Happiness* into *Positive*; relabelled the labels *Disgust* and *Repression* into *Negative*; maintained the label *Surprise* with *Surprise*. Tab. I summarize the essential information.

Protocol. The cross-database protocol is designed to develop models with promising domain adaption performance operated by training model in the Source database (S) and testing in the Target database (T), which is denoted as S→T. Following [28], we employed two types of unsupervised CDMER experiments: TYPE-I is implemented between every two subsets of SMIC, and TYPE-II is implemented between Selected CASME II and any subset of SMIC. We denote SMIC-HS, SMIC-VIS, and SMIC-NIR as H, V, N, and CASME II as C for short. Specially, TYPE-I experiment includes six experiments: H→V, V→H, H→N, N→H, V→N, N→V, TYPE-II experiment consists of another six experiments: C→H, H→C, C→V, V→C, C→N, N→C.

Evaluation Metrics. We employed macro F1-score (M-F1) and accuracy (ACC) to evaluate our method. Macro F1-score is calculated by $M-F1=\frac{1}{C}\sum_{c=1}^{C}\frac{2p_c r_c}{p_c+r_c}$, where p_c and r_c are the precision and recall of the c-th category micro-expression, and C is the category number. M-F1 is appropriate because the unbalanced sample problem widely existed in the CDMER.

Implementation Detail. In the experiments, we constructed the bounding box for face cropping using the facial landmarks from the first frame of ME sequence. We employed the Temporal Interpolation Model (TIM)[40], [41] to convert the ME sequence into fixed 16 frames in temporal and resized each frame into 112×112 pixels in spatial. For each ME sequence, we used a grid-based multi-scale spatial division scheme to divide the whole face into four scales of 1×1, 2×2, 4×4, 8×8, a total of $K = 85$ local face sequences, then extracted and concatenated the corresponding LBP-TOP features[18] to serve as the micro-expression representation. Here, the neighboring radius of LBP-TOP and the number of neighboring points are set to $R = 3$ and $P - 8$. Two hyper-parameters are involved in solving our proposed method, i.e., the salient facial region number κ and the weighting hyper-parameter ξ of the MMD term. Here, salient facial block number κ is an integer variable, and weighting hyper-parameter ξ is a consecutive variable. Following the work of [38], [28], we used a grid-based searching strategy to search the optimal hyper-parameters for achieving the best M-F1 performance, and reported the M-F1 and ACC metrics under the corresponding setting. Specially, we searched the hyper-parameter κ from a preset parameter interval [1:1:85], and searched the hyper-parameter ξ from a preset parameter interval [0.001:0.0002:0.01 0.01:0.002:0.1 0.1:0.02:1 1:0.2:10 10:2:100 100:20:1000].

B. Results and Analysis

Tab. II and Tab. III show TYPE-I and TYPE-II CDMER experiments respectively. We calculated and listed the average M-F1 and ACC performance of these two types of experiments. In the TYPE-I experiments (Exp.1 to Exp.6), SVM[31], IW-SVM[32], TCA[33], GFK[34], SA[35], STM[36], TKL[37], TSRG[38], DRLS[39] achieved 0.6003, 0.6911, 0.6238, 0.7223, 0.6917, 0.6440, 0.6964, 0.6991, 0.6552 on the M-F1 metrics, and achieved 61.62%, 69.74%, 64.01%, 72.44%, 69.44%, 65.78%, 69.92%, 70.05%, 66.60% on the ACC metrics. In the TYPE-II experiments (Exp.7 to Exp.12), SVM[31], IW-SVM[32], TCA[33], GFK[34], SA[35], STM[36], TKL[37], TSRG[38], DRLS[39] achieved 0.4112, 0.4876, 0.5177, 0.5161, 0.4877, 0.3982, 0.5051, 0.5348, 0.5102 on the M-F1 metrics, and achieved 45.55%, 50.73%, 53.45%, 54.31%, 50.90%, 45.61%, 52.33%, 56.22%, 53.71% on the ACC metrics. We show the best result for each experiment in Tab. II and Tab. III in bold. From the bold-lighted in Tab. II and Tab. III, we observed that proposed TGSR outperforms those state-of-the-art methods beyond half experiments and achieves the best performance in 7 of the total 12 CDMER experiments. In the TYPE-I experiments, from Exp.1 to Exp.6, the best M-F1 is achieved at the hyper-parameters (κ,ξ) value of (85, 0.0022), (46, 0.0036), (14, 4000), (85, 0.0044), (12, 44), (12, 280), respectively. In the TYPE-II experiments, from Exp.7 to Exp.12, the best M-F1 is achieved at the hyper-parameters (κ,ξ) value of (62,0.0012), (28,0.0980), (85,0.0030), (85,0.0280), (85,0.0016), (75,0.0220), respectively.

In addition, at least three apparent characteristics we can find in Tab. II and Tab. III. Firstly, all comparison methods except for SVM use domain adaption techniques, and consistently outperform SVM on M-F1 and ACC metrics in two types of experiments. Taking IW-SVM for example, it effectively improves SVM by learning a group of domain adaptive weighting coefficient: model performance improved 0.0908 and 8.12% on average M-F1 and ACC metrics in TYPE-I experiments, and improved 0.0764 and 5.18% on average M-F1 and ACC metrics in TYPE-II experiments. Thus we infer that domain adaptive techniques can improve model

TABLE II

THE RESULTS OF TYPE-I CDMER EXPERIMENTS ARE BASED ON ANY TWO SUBSETS OF SMIC, I.E., SMIC-HS, SMIC-VIS, AND SMIC-NIR. THE MICRO-EXPRESSION CATEGORY INCLUDES *Negative*, *Positive*, AND *Surprise*. THE BEST RESULTS FROM EACH EXPERIMENT ARE SHOWN IN BOLD. WE USE MACRO F1-SCORE (M-F1) AND ACCURACY (ACC) TO EVALUATE METHODS.

Method	Exp.1(H→V)		Exp.2(V→H)		Exp.3(H→N)		Exp.4(N→H)		Exp.5(V→N)		Exp.6(N→V)	
	M-F1	ACC	M-F1	ACC	M-F1	ACC	M-F1	ACC	M-F1	ACC	M-F1	ACC
SVM[31]	0.8002	80.28	0.5421	54.27	0.5455	53.52	0.4878	54.88	0.6186	63.38	0.6078	63.38
IW-SVM[32]	0.8868	88.73	0.5852	58.54	**0.7469**	**74.65**	0.5427	54.27	0.6620	69.01	0.7228	73.24
TCA[33]	0.8269	83.10	0.5477	54.88	0.5828	59.15	0.5443	57.32	0.5810	61.97	0.6598	67.61
GFK[34]	0.8448	84.51	0.5957	59.15	0.6977	70.42	0.6197	62.80	**0.7619**	**76.06**	0.8142	81.69
SA[35]	0.8037	80.28	0.5955	59.15	0.7465	**74.65**	0.5644	56.10	0.7004	71.83	0.7394	74.65
STM[36]	0.8253	83.10	0.5059	51.22	0.6628	66.20	0.5351	56.10	0.6427	67.61	0.6922	70.42
TKL[37]	0.7742	77.46	0.5738	57.32	0.7051	70.42	0.6116	62.20	0.7558	76.06	0.7580	76.06
TSRG[38]	0.8869	88.73	0.5652	56.71	0.6484	64.79	0.5770	57.93	0.7056	70.42	0.8116	81.69
DRLS[39]	0.8604	85.92	0.6120	60.98	0.6599	66.20	0.5599	55.49	0.6620	69.01	0.5771	61.97
Ours	**0.9150**	**91.55**	**0.6226**	**62.20**	0.5847	60.56	**0.6272**	**61.59**	0.6984	70.42	**0.8403**	**84.51**

TABLE III

THE RESULTS OF TYPE-II CDMER EXPERIMENTS ARE BASED ON SELECTED CASME II DATABASE AND ANY ONE SUBSET OF SMIC DATABASES, I.E., ONE OF SMIC-HS, SMIC-VIS, AND SMIC-NIR. THE MICRO-EXPRESSION CATEGORY INCLUDES *Negative*, *Positive*, AND *Surprise*. THE BEST RESULTS FROM EACH EXPERIMENT ARE SHOWN IN BOLD. WE USE MACRO F1-SCORE (M-F1) AND ACCURACY (ACC) TO EVALUATE METHODS.

Method	Exp.7(C→H)		Exp.8(H→C)		Exp.9(C→V)		Exp.10(V→C)		Exp.11(C→N)		Exp.12(N→C)	
	M-F1	ACC	M-F1	ACC	M-F1	ACC	M-F1	ACC	M-F1	ACC	M-F1	ACC
SVM[31]	0.3697	45.12	0.3245	48.46	0.4701	50.70	0.5367	53.08	0.5295	52.11	0.2368	23.85
IW-SVM[32]	0.3541	41.46	0.5829	62.31	0.5778	59.15	0.5537	54.62	0.5117	50.70	0.3456	36.15
TCA[33]	0.4637	46.34	0.4870	53.08	**0.6834**	**69.01**	0.5789	59.23	0.4992	50.70	0.3937	42.31
GFK[34]	0.4126	46.95	0.4776	50.77	0.6361	66.20	0.6056	61.50	0.5180	53.52	0.4469	46.92
SA[35]	0.4302	47.56	0.5447	62.31	0.5939	59.15	0.5243	51.54	0.4738	47.89	0.3592	36.92
STM[36]	0.3640	43.90	**0.6115**	**63.85**	0.4051	52.11	0.2715	30.00	0.3523	42.25	0.3850	41.54
TKL[37]	0.4582	46.95	0.4661	54.62	0.6042	60.56	0.5378	53.08	0.5392	54.93	0.4248	43.85
TSRG[38]	**0.5042**	51.83	0.5171	60.77	0.5935	59.15	0.6208	63.08	0.5624	56.34	0.4105	46.15
DRLS[39]	0.4924	**53.05**	0.5267	59.23	0.5757	57.75	0.5942	60.00	0.4885	49.83	0.3838	42.37
Ours	0.5001	51.83	0.5061	56.92	0.5906	59.15	**0.6403**	**63.85**	**0.5697**	**57.75**	**0.4474**	**48.46**

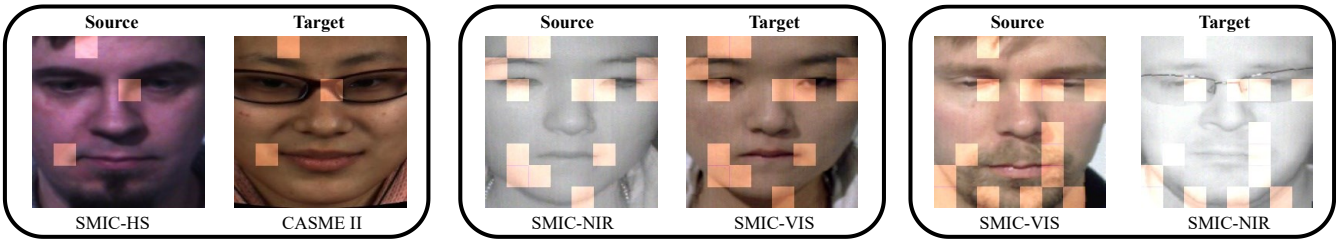

Fig. 2. The salient facial regions selected by our proposed TGSR method in three cross-database micro-expression recognition task. The sub-matrix $\hat{C}_i$ corresponding to the salient regions is the matrix filled with scalar one.

performance on CDMER tasks. Secondly, we observe that the results of TYPE-I experiments are generally better than TYPE-II experiments. We believe the experimental setup itself caused it. The TYPE-I experiments selected two subsets of SMIC database with different imaging modalities and TYPE-II experiments used Selected CASME II and one subset of SMIC database, as the source and target databases respectively. It is clear that the database differences of TYPE-I experiments are more significant than TPYE-II experiments. Thirdly, we find that a noticeable performance gap existed in those experiments exchanging the source and target databases: all performance on Exp.1(H→V) are generally better than those on Exp.2(V $Rightarrow$H); all performance on Exp.3(H→N) are generally better than those on Exp.4(N→H); all performance on Exp.11(C→N) are generally better than those on Exp.12(N→C). Exp.1(H→V) used the a high-speed camera captured image sequences from the SMIC-HS subset as the source database and the general visual camera captured image sequence from the SMIC-VIS subset as the target database, which is exchanged in Exp.2(V→H). Exp.3(H→V) used colored image sequences captured by high-speed camera from SMIC-HS subset as the source database and the uncolored near-infrared image sequence from SMIC-NIR subset as the target database, which is exchanged in Exp.4(N→H).

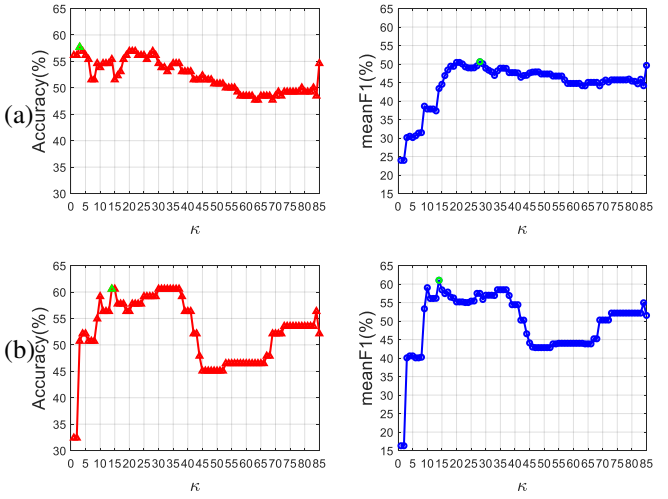

Fig. 3. The performance curve of the proposed TGSR method under different hyper-parameter κ, i.e., the salient facial region number. (a) shows the experimental results of Exp.8(H→C) and (b) shows the experimental results of Exp.4(H→N).

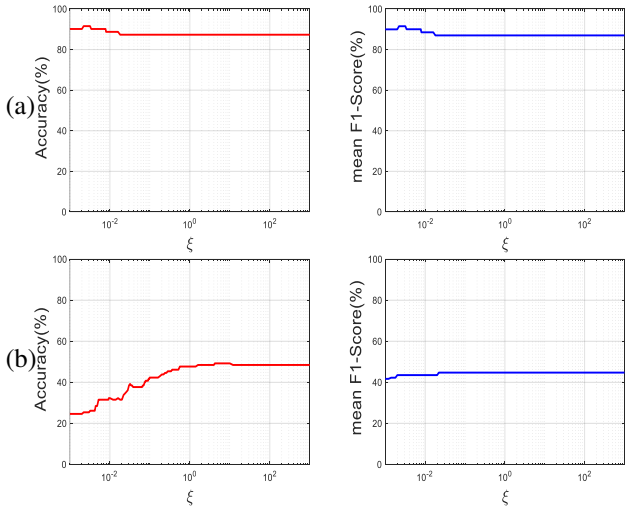

Fig. 4. The performance curve of the proposed TGSR method under different hyper-parameter ξ, i.e., weighting hyper-parameter of the MMD term. (a) shows the experimental results of Exp.1(H→V) and (b) shows the experimental results of Exp.12(N→C).

Exp.11(C→N) used colored image sequences from Selected CASME II database as the source database and uncolored image sequences from SMIC-NIR subset as the target database, which is exchanged in Exp.12(N→C). We believe the reason why the performances on Exp.1(H→V) are generally better than Exp.2(V→H) and the performances on Exp.3(H→N) are generally better than Exp.4(N→H) is that high-speed camera can capture more subtle facial movements And the reason why the performances on Exp.11(C→N) are generally better than Exp.12(N→C) is that the Selected CASME II database retains but the SMIC-NIR subset discards color information such crucial cue for understanding human facial expressions[42]. In addition, we observe that although Exp.3(N→H)-Exp.4(N→H) pair and Exp.11(C→N)-Exp.12(N→C) pair both have a color difference between the source and target databases, the model performance gap between Exp.11(C→N)-Exp.12(N→C) pair is more significant than Exp.3(N→H)-Exp.4(N→H) pair. It may be due to other database differences other than the color.

C. Hyper-parameter Discussion & Visualization

Two hyper-parameters are involved in solving the optimal regression matrix $\hat{C}$ of proposed TGSR, i.e., the salient facial region number κ and the MMD weighting hyper-parameter ξ. The setting of these hyper-parameters affects model performance, thus we conducted two experiments to investigate model sensitiveness to hyper-parameters κ and ξ. In the first experiment, we fixed hyper-parameter ξ and varied hyper-parameter κ from 1 to $K = 85$, then recorded corresponding M-F1 and ACC metrics, to explore model the sensitiveness to hyper-parameter κ. We selected Exp.4(H→N) and Exp.8(H→C) as the typical of TYPE-I and TYPE-II CDMER experiments respectively, and presented their performance curve as Fig. 3 shown. We can see that the M-F1 and ACC performance of TGSR model increases with hyper-parameter κ increases firstly, and reaches its peak at a low κ value, then decreases with hyper-parameter κ increases. It means that the salient facial regions for CDMER are exiguous.In the second experiment, we fixed hyper-parameter κ and varied hyper-parameter ξ from 10^{-3} to 10^{3}, then recorded corresponding M-F1 and ACC metrics, to explore the model sensitiveness to hyper-parameter ξ. We selected Exp.1(H→V) and Exp.12(N→C) as the typical of TYPE-I and TYPE-II CDMER experiments respectively, and presented their performance curve as Fig. 4 shown. It is apparent that selecting an appropriate value of weighting hyper-parameter ξ helps TGSR yields better performance, and the MMD term can effectively and stably improve model performance across a wide range of hyper-parameter ξ. We also selected Exp.5(V→N), Exp.6(N→V), and Exp.8(H→C) as the typical to visualize the learned salient facial regions for CDMER. From Fig. 2, we observed that the selected facial regions are consistent with micro-expression's AU definition, thus we can believe that our proposed TGSR achieved a competitive performance by learning an explicable feature.

IV. Conclusion

This paper proposes a novel Transfer Group Sparse Regression, TGSR, to cope with the Cross-Database Micro-Expression Recognition (CDMER) problem. TGSR selects the salient facial regions by learning a binary regression matrix shared between the source and target databases and performing sparse operations at the feature level to 1) optimize the difference between the source and target databases so as to alleviate this difference better, 2) highlight the valid part to make extracted feature more effective and explicable. Experiments and visualizations show that TGSR learns explicable and highly discriminative micro-expression features and outperforms most state-of-the-art subspace-learning-based domain-adaptive methods for CDMER.

ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China (NSFC) under the Grants U2003207, 61921004, and 61902064, in part by the Fundamental Research Funds for the Central Universities under Grant 2242022k30036, and in part by the Zhishan Young Scholarship of Southeast University.

REFERENCES

[1] P. Ekman and W. V. Friesen, "Nonverbal leakage and clues to deception," *Psychiatry*, vol. 32, no. 1, pp. 88–106, 1969. 1

[2] P. Ekman, *Telling lies: Clues to deceit in the marketplace, politics, and marriage (revised edition)*. WW Norton & Company, 2009. 1

[3] W.-J. Yan, Q. Wu, J. Liang, Y.-H. Chen, and X. Fu, "How fast are the leaked facial expressions: The duration of micro-expressions," *Journal of Nonverbal Behavior*, vol. 37, no. 4, pp. 217–230, 2013. 1

[4] M. Frank, M. Herbasz, K. Sinuk, A. Keller, and C. Nolan, "I see how you feel: Training laypeople and professionals to recognize fleeting emotions," in *The Annual Meeting of the International Communication Association. Sheraton New York, New York City*, 2009, pp. 1–35. 1

[5] X. Jiang, Y. Zong, W. Zheng, C. Tang, W. Xia, C. Lu, and J. Liu, "Dfew: A large-scale database for recognizing dynamic facial expressions in the wild," in *Proc. ACM MM*, 2020, pp. 2881–2889. 1

[6] S. Li, W. Zheng, Y. Zong, C. Lu, C. Tang, X. Jiang, J. Liu, and W. Xia, "Bi-modality fusion for emotion recognition in the wild," in *Proc. ICMI*, 2019, pp. 589–594. 1

[7] C. Yang, H. Wu, Z. Li, W. He, N. Wang, and C.-Y. Su, "Mind control of a robotic arm with visual fusion technology," *IEEE Trans. Ind. Informat.*, vol. 14, no. 9, pp. 3822–3830, 2017. 1

[8] C. Yang, J. Luo, Y. Pan, Z. Liu, and C.-Y. Su, "Personalized variable gain control with tremor attenuation for robot teleoperation," *IEEE Trans. Syst., Man, Cybern.*, vol. 48, no. 10, pp. 1759–1770, 2017. 1

[9] T. Pfister, X. Li, G. Zhao, and M. Pietikäinen, "Recognising spontaneous facial micro-expressions," in *Proc. ICCV*. IEEE, 2011, pp. 1449–1456. 1

[10] Y. Wang, J. See, R. C.-W. Phan, and Y.-H. Oh, "Lbp with six intersection points: Reducing redundant information in lbp-top for micro-expression recognition," in *Proc. ACCV*. Springer, 2014, pp. 525–537. 1

[11] S.-J. Wang, W.-J. Yan, X. Li, G. Zhao, C.-G. Zhou, X. Fu, M. Yang, and J. Tao, "Micro-expression recognition using color spaces," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 6034–6047, 2015. 1

[12] Y.-J. Liu, J.-K. Zhang, W.-J. Yan, S.-J. Wang, G. Zhao, and X. Fu, "A main directional mean optical flow feature for spontaneous micro-expression recognition," *IEEE Trans. Affect. Comput.*, vol. 7, no. 4, pp. 299–310, 2015. 1

[13] D. H. Kim, W. J. Baddar, and Y. M. Ro, "Micro-expression recognition with expression-state constrained spatio-temporal feature representations," in *Proc. ACM MM*, 2016, pp. 382–386. 1

[14] P. Lu, W. Zheng, Z. Wang, Q. Li, Y. Zong, M. Xin, and L. Wu, "Micro-expression recognition by regression model and group sparse spatio-temporal feature learning," *IEICE Transactions on Information and Systems*, vol. 99, no. 6, pp. 1694–1697, 2016. 1

[15] F. Xu, J. Zhang, and J. Z. Wang, "Microexpression identification and categorization using a facial dynamics map," *IEEE Trans. Affect. Comput.*, vol. 8, no. 2, pp. 254–267, 2017. 1

[16] S. Happy and A. Routray, "Fuzzy histogram of optical flow orientations for micro-expression recognition," *IEEE Trans. Affect. Comput.*, vol. 10, no. 3, pp. 394–406, 2017. 1

[17] Y. Zong, X. Huang, W. Zheng, Z. Cui, and G. Zhao, "Learning from hierarchical spatiotemporal descriptors for micro-expression recognition," *IEEE Trans. Multimedia*, vol. 20, no. 11, pp. 3160–3172, 2018. 1

[18] 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, 2007. 1, 4

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Proc. NeurIPS*, vol. 25, pp. 1097–1105, 2012. 1

[20] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997. 1

[21] M. Wei, W. Zheng, Y. Zong, X. Jiang, C. Lu, and J. Liu, "A novel micro-expression recognition approach using attention-based magnification-adaptive networks," in *Proc. ICASSP*. IEEE, 2022, pp. 2420–2424. 1

[22] W. Xia, W. Zheng, Y. Zong, and X. Jiang, "Motion attention deep transfer network for cross-database micro-expression recognition," in *Proc. ICPR*. Springer, 2021, pp. 679–693. 1

[23] Y. Zong, W. Zheng, Z. Cui, G. Zhao, and B. Hu, "Toward bridging microexpressions from different domains," *IEEE Transactions on Cybernetics*, vol. 50, no. 12, pp. 5047–5060, 2019. 1

[24] L. Li, X. Zhou, Y. Zong, W. Zheng, X. Chen, J. Shi, and P. Song, "Unsupervised cross-database micro-expression recognition using target-adapted least-squares regression," *IEICE Transactions on Information and Systems*, vol. 102, no. 7, pp. 1417–1421, 2019. 1

[25] X. Chen, X. Zhou, C. Lu, Y. Zong, W. Zheng, and C. Tang, "Target-adapted subspace learning for cross-corpus speech emotion recognition," *IEICE Transactions on Information and Systems*, vol. 102, no. 12, pp. 2632–2636, 2019. 1

[26] W.-J. Yan, X. Li, S.-J. Wang, G. Zhao, Y.-J. Liu, Y.-H. Chen, and X. Fu, "Casme ii: An improved spontaneous micro-expression database and the baseline evaluation," *PLOS One*, vol. 9, no. 1, p. e86041, 2014. 1, 3

[27] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikäinen, "A spontaneous micro-expression database: Inducement, collection and baseline," in *Proc. FG*. IEEE, 2013, pp. 1–6. 1, 4

[28] T. Zhang, Y. Zong, W. Zheng, C. P. Chen, X. Hong, C. Tang, Z. Cui, and G. Zhao, "Cross-database micro-expression recognition: A benchmark," *IEEE Trans. Knowl. Data Eng.*, 2020. 1, 4

[29] Z. T. Qin and D. Goldfarb, "Structured sparsity via alternating direction methods." *Journal of Machine Learning Research*, vol. 13, no. 5, 2012. 2

[30] Z. Lin, M. Chen, and Y. Ma, "The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices," *arXiv preprint arXiv:1009.5055*, 2010. 2

[31] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 1–27, 2011. 4, 5

[32] A. Hassan, R. Damper, and M. Niranjan, "On acoustic emotion recognition: compensating for covariate shift," *IEEE Trans. Speech Audio Process.*, vol. 21, no. 7, pp. 1458–1468, 2013. 4, 5

[33] S. J. 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, 2010. 4, 5

[34] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in *Proc. CVPR*. IEEE, 2012, pp. 2066–2073. 4, 5

[35] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars, "Unsupervised visual domain adaptation using subspace alignment," in *Proc. ICCV*, 2013, pp. 2960–2967. 4, 5

[36] W.-S. Chu, F. De la Torre, and J. F. Cohn, "Selective transfer machine for personalized facial action unit detection," in *Proc. CVPR*, 2013, pp. 3515–3522. 4, 5

[37] M. Long, J. Wang, J. Sun, and S. Y. Philip, "Domain invariant transfer kernel learning," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 6, pp. 1519–1532, 2014. 4, 5

[38] Y. Zong, X. Huang, W. Zheng, Z. Cui, and G. Zhao, "Learning a target sample re-generator for cross-database micro-expression recognition," in *Proc. ACM MM*, 2017, pp. 872–880. 4, 5

[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, 2018. 4, 5

[40] Z. Zhou, G. Zhao, and M. Pietikäinen, "Towards a practical lipreading system," in *Proc. CVPR*. IEEE, 2011, pp. 137–144. 4

[41] Z. Zhou, X. Hong, G. Zhao, and M. Pietikäinen, "A compact representation of visual speech data using latent variables," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 1, pp. 1–1, 2013. 4

[42] C. F. Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, "Facial color is an efficient mechanism to visually transmit emotion," *Proceedings of the National Academy of Sciences*, vol. 115, no. 14, pp. 3581–3586, 2018. 6