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 aims to 1) optimize the measurement and better alleviate the difference between the source and target databases, and 2) highlight the valid facial regions to enhance extracted features, by the operation of selecting the group features from the raw face feature, where each region is associated with a group of raw face feature, i.e., the salient facial region selection. Compared with previous transfer group sparse methods, our proposed TGSR has the ability to select the salient facial regions, which is effective in alleviating aforementioned problems for better performance and reducing the computational cost at the same time. We use two public ME databases, i.e., CASME II and SMIC, to evaluate our proposed TGSR method. Experimental results show that our proposed TGSR learns the discriminative and explicable regions, 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], etc.

Due to the huge potential value of MEs, many efforts have been made to design an automatic Micro-Expression Recognition (MER) system over the last few decades[7], [8], [9], [10], [11]. They developed many subspace-learning-based methods[12], [7], [13] and deep learning methods[14], [15], and promoted the rapid development of automated MER technology. However, most existing methods are evaluated on a single database, which may sharply drop the performance when applied in domains different from the training database, such as imaging devices, subjects, scenes, etc.

To learn a domain-robust MER model, researchers have turned their interests to the domain-adaptive MER method recently. A new challenging topic has thus emerged, i.e., Cross-Database Micro-Expression Recognition (CDMER). It mimics the domain variation problem and evaluates the method's adaptive ability by the operation of training the model in one micro-expression database, i.e., the source database, and testing in the other one, i.e., the target database. CDMER[16], [17], [18] 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. Past research has proposed many subspace-learning-based methods to yield a similar and effective feature to bridge this gap between the source and target domains and effectively advance the MER model's adaptive ability. However, existing CDMER models' performance is still far from satisfactory.

Salient facial region is an approach to enhancing the few but discriminative regions and suppressing the many but noisy regions, aiming to improve performance and reduce computational cost simultaneously. It has been widely validated that the salient region selection approach benefits emotion recognition performance. Inspired by this, we introduce a learnable binary sparse regression matrix shared between the source and target databases, and propose a novel Transfer Group Sparse Regression method (TGSR) to cope with the CDMER problem with the assistance of salient facial region selection technology. Our proposed TGSR contains three terms: a regression term with the learnable matrix for bridging micro-expression features and labels, a joint feature distribution regularization term for measuring and alleviating the difference between source and target databases, and a regression matrix sparse term to promote our proposed TGSR learns the few but discriminative region feature. Especially, the salient facial region selection in our proposed TGSR is achieved by the operation of selecting the group features from the raw features, where each region is associated with a group of raw face features. And the facial region selection is aimed to seek the discriminative regions for 1) optimizing the measurement and better alleviating the difference between the source and target databases, and 2) highlighting the valid facial regions to enhance extracted features, which will significantly improve the performance of CDMER model. In addition, facial region selection can also reduce the computational cost when pursuing better CDMER performance. We evaluate our method on CASME II[19] and SMIC[20] databases. Experimental results and corresponding visualization show that our proposed TGSR can seek the salient and explicable facial regions to alleviate the afore-

# indicates the corresponding authors

mentioned problems effectively and outperform most state-of-the-art subspace-learning-based domain-adaptive methods for CDMER.

II. Method

A. The Generation of Micro-Expression Features

Extracting facial features is the first step for CDMER. As Fig. 1 shown, we firstly use the grid-based multi-scale spatial division scheme[21] to divide the cropped ME sequence into four scales in total of K regions, i.e., K spatial local sequences. Then we extracted d-dimensional feature $\boldsymbol{x}_k$ of K facial region, $k \in [1, K]$, and obtain sample's multi-scale hierarchical feature $\boldsymbol{x}^{\nu} = \left[\boldsymbol{x}_1^{\mathrm{T}}, \cdots, \boldsymbol{x}_K^{\mathrm{T}}\right]^{\mathrm{T}} \in \mathbb{R}^{Kd}$ by concatenating region features one by one. Suppose that we have N_s source and N_t target micro-expression samples, the feature matrix of the source and target databases 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 $\boldsymbol{X}^t = \left[\boldsymbol{X}_1^{t\,\mathrm{T}}, \cdots, \boldsymbol{X}_K^{t\,\mathrm{T}}\right]^{\mathrm{T}} \in \mathbb{R}^{Kd \times N_t}$, respectively. Here, each column of $\boldsymbol{X}^s$ and $\boldsymbol{X}^t$ is a feature vector like $\boldsymbol{x}^{\nu}$, they respectively denote the feature of single micro-expression sample from the corresponding 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 group feature corresponding to the i-th facial region from the source and target databases. The labels 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.

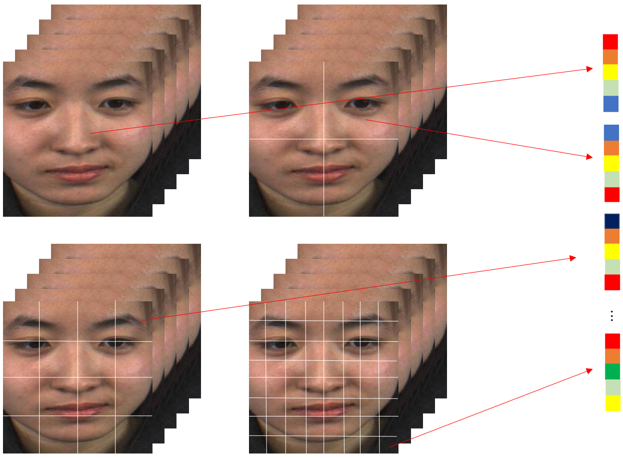

Fig. 1. The grid-based multi-scale spatial division scheme for extracting micro-expression features.

B. Proposed Method

The basic idea of our proposed Transfer Group Sparse Regression method (TGSR) is improve the CDMER performance by sailent facial region selection approach, i.e., selecting the group features from the raw face feature, where each region is associated with a group of raw face feature. Our proposed TGSR contains three terms: 1) a regression term with the learnable regression matrix for bridging micro-expression features and labels, 2) a joint feature distribution regularization term for measuring and alleviating the difference between source and target databases, and 3) a regression matrix sparse term to promote our proposed TGSR learns the few but discriminative region feature, which can be denoted 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 $\boldsymbol{C} = [\boldsymbol{C}_1^{\mathrm{T}}, ..., \boldsymbol{C}_K^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{Kd \times C}$ is such a domain-invariant regression matrix, the sub-matrix $\boldsymbol{C}_i$ of $\boldsymbol{C}$ is to construct the relation between the group feature of i-th facial region and corresponding sample labels, $f_1(\boldsymbol{C}_i)$ is the joint feature distribution regularization term, $f_2(\boldsymbol{C}_i)$ is the regression matrix sparse term, and ξ and λ the weighting hyper-parameters of $f_1(\boldsymbol{C}_i)$ and $f_2(\boldsymbol{C}_i)$.

By minimizing $f_1(\boldsymbol{C}_i)$ together with the regression term, we can alleviate the database difference. We use the maximum mean discrepancy (MMD) to serve as this regularization term, which can be expressed as Equ. (2),

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

where $\varPhi(\cdot)$ 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 features 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 difference in label space, Equ. (2) becomes solvable. $f_2(\boldsymbol{C}_i)$ is defined as Equ. (4). Our proposed TGSR is promoted to select the few but discriminative regions when minimizing $f_2(\boldsymbol{C}_i)$ together with the regression matrix.

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

By substituting Equ. (3) and Equ. (4) into Equ. (1), we can rewrite the objective function of Equ. (1) 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)[22] and Inexact Augmented Lagrangian Multiplier (IALM)[23] to solve Equ. (5) involving facial region selection technology, i.e., selecting the group features of the raw face features, under presetting salient region number.

We firstly rewrite Equ. (5) into Equ. (6),

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

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

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

where $\boldsymbol{P}_i \in \mathbb{R}^{d\times C}$ denotes the Lagrangian multiplier matrix corresponding to the i-th facial region, 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 until convergence, and *Algorithm 1* show more details:

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),

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

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),

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

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

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

3) Update $\boldsymbol{P}$ and μ.

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

Algorithm 1 The Algorithm for solving the optimal regression matrix $\boldsymbol{C}$ in our proposed 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 $\kappa$, the scalar parameter $\rho$, $\mu_{max}$.
  • 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 $\mu$.
Repeating steps 1) to 4) until convergence.
1: Fix $\boldsymbol{C}, \boldsymbol{P}, \mu$ and update $\boldsymbol{D}$:
       $\boldsymbol{D} = \left(\mu\boldsymbol{I}_{Kd} + 2\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^{\mathrm{T}}\right)^{-1}\left(2\tilde{\boldsymbol{X}}\tilde{\boldsymbol{L}}^{\mathrm{T}} + \boldsymbol{P} + \mu\boldsymbol{C}\right)$,   (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
       $\boldsymbol{C}_i = \begin{cases} \frac{d_i - \frac{\lambda}{\mu}}{d_i}(\boldsymbol{D}_i - \frac{\boldsymbol{P}_i}{\mu}), & \frac{\lambda}{\mu} < d_i, \\ \boldsymbol{0}, & \frac{\lambda}{\mu} \geq d_i. \end{cases}$   (13)
3: Update $\boldsymbol{P}$ and $\mu$:
       $\boldsymbol{P} = \boldsymbol{P} + \mu\left(\boldsymbol{D} - \boldsymbol{C}\right)$, $\mu = \min\left(\rho\mu, \mu_{max}\right)$
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 approach. Then, we can extract the feature $\boldsymbol{x}_i^{te} \in \mathbb{R}^{Kd}$ of the micro-expression sample to be predicted and estimate 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 this micro-expression sample to the largest entry index of the predicted label vector, i.e., micro-expression category $\hat{c}$.

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

III. EXPERIMENT

A. Experiment Setup

1) Database: We evaluated our method on Selected CASME II and SMIC databases. CASME II[19] 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 of *Disgust*, *Happiness*, *Repression*, and *Surprise* to be the Selected CASME II. SMIC[20] 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.

2) Protocol: The cross-database protocol is designed to develop models with promising domain adaption performance operated by training the model in the Source database (S) and testing in the Target database (T), which is denoted as S→T. Following [21], 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.

3) 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.

4) Data pre-processing: We firstly cropped the whole face of each ME sequence using the bounding box from the first frame. Then we employed the Temporal Interpolation Model (TIM)[33], [34] to convert the ME sequence into fixed 16 frames in temporal. And resized each frame into 112 × 112 pixels in spatial.

5) Feature extraction: 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, i.e., facial regions. Then extracted and concatenated the corresponding LBP-TOP features[12] of these facial regions 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 $P = 8$.

6) Training setting: Two hyper-parameters are involved in solving our proposed TGSR, i.e., the salient facial region number κ and the weighting hyper-parameter ξ of the MMD term. Following the work of [31], [21], we used a grid-based searching strategy to search the optimal hyper-parameters of our proposed TGSR for achieving the best M-F1 performance. We reported both M-F1 and ACC metrics under the optimal 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

1) Overall results: We bold-lighted the best result of each experiment in Tab. II and Tab. III. We observed that our proposed TGSR outperforms those state-of-the-art methods beyond half experiments and achieved the best performance in 7 of total 12 CDMER experiments. It indicates that our proposed TGSR has the ability to cope with the CDMER problem effectively. We also reported the hyper-parameters to achieve this performance. 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. The best performance tends to select the few but discriminative regions rather than all facial regions. And our proposed TGSR also displays competitive performance on those experiments that do not work best. It also indicates the effectiveness of our salient facial region selection strategy.

2) Difference analysis: Two apparent performance characteristics involving database difference can be found in Tab. II and Tab. III.

Firstly, 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

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)		Average	
	M-F1	ACC	M-F1	ACC	M-F1	ACC	M-F1	ACC	M-F1	ACC	M-F1	ACC	M-F1	ACC
SVM[24]	0.8002	80.28	0.5421	54.27	0.5455	53.52	0.4878	54.88	0.6186	63.38	0.6078	63.38	0.6003	61.62
IW-SVM[25]	0.8868	88.73	0.5852	58.54	**0.7469**	**74.65**	0.5427	54.27	0.6620	69.01	0.7228	73.24	0.6911	67.74
TCA[26]	0.8269	83.10	0.5477	54.88	0.5828	59.15	0.5443	57.32	0.5810	61.97	0.6598	67.61	0.6238	64.01
GFK[27]	0.8448	84.51	0.5957	59.15	0.6977	70.42	0.6197	**62.80**	**0.7619**	**76.06**	0.8142	81.69	**0.7223**	**72.44**
SA[28]	0.8037	80.28	0.5955	59.15	0.7465	**74.65**	0.5644	56.10	0.7004	71.83	0.7394	74.65	0.6917	69.44
STM[29]	0.8253	83.10	0.5059	51.22	0.6628	66.20	0.5351	56.10	0.6427	67.61	0.6922	70.42	0.6440	65.78
TKL[30]	0.7742	77.46	0.5738	57.32	0.7051	70.42	0.6116	62.20	0.7558	76.06	0.7580	76.06	0.6964	69.92
TSRG[31]	0.8869	88.73	0.5652	56.71	0.6484	64.79	0.5770	57.93	0.7056	70.42	0.8116	81.69	0.6991	70.05
DRLS[32]	0.8604	85.92	0.6120	60.98	0.6599	66.20	0.5599	55.49	0.6620	69.01	0.5771	61.97	0.6552	66.60
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**	0.7141	71.80

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)		Average	
	M-F1	ACC	M-F1	ACC	M-F1	ACC	M-F1	ACC	M-F1	ACC	M-F1	ACC	M-F1	ACC
SVM[24]	0.3697	45.12	0.3245	48.46	0.4701	50.70	0.5367	53.08	0.5295	52.11	0.2368	23.85	0.4112	45.55
IW-SVM[25]	0.3541	41.46	0.5829	62.31	0.5778	59.15	0.5537	54.62	0.5117	50.70	0.3456	36.15	0.4876	50.73
TCA[26]	0.4637	46.34	0.4870	53.08	**0.6834**	**69.01**	0.5789	59.23	0.4992	50.70	0.3937	42.31	0.5177	53.45
GFK[27]	0.4126	46.95	0.4776	50.77	0.6361	66.20	0.6056	61.50	0.5180	53.52	0.4469	46.92	0.5161	54.31
SA[28]	0.4302	47.56	0.5447	62.31	0.5939	59.15	0.5243	51.54	0.4738	47.89	0.3592	36.92	0.4877	50.90
STM[29]	0.3640	43.90	**0.6115**	**63.85**	0.4051	52.11	0.2715	30.00	0.3523	42.25	0.3850	41.54	0.3982	45.61
TKL[30]	0.4582	46.95	0.4661	54.62	0.6042	60.56	0.5378	53.08	0.5392	54.93	0.4248	43.85	0.5051	52.33
TSRG[31]	**0.5042**	51.83	0.5171	60.77	0.5935	59.15	0.6208	63.08	0.5624	56.34	0.4105	46.15	0.5348	56.22
DRLS[32]	0.4924	**53.05**	0.5267	59.23	0.5757	57.75	0.5942	60.00	0.4885	49.83	0.3838	42.37	0.5102	53.71
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**	**0.5424**	**56.33**

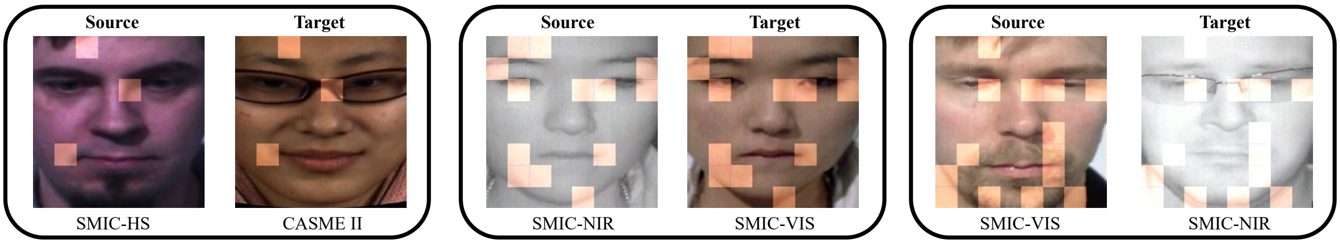

Fig. 2. The salient facial regions selected by our proposed TGSR method in three cross-database micro-expression recognition tasks. The sub-matrix $\hat{C}_i$ corresponding to the salient regions is the matrix filled with scalar one.

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.

Secondly, 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→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→N) 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). 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

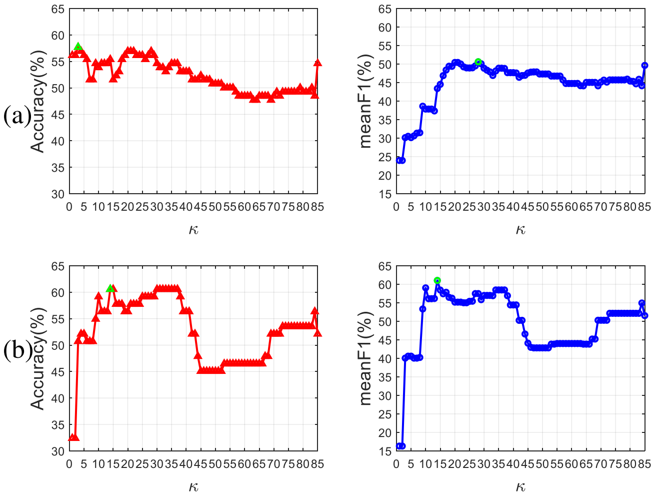

Fig. 3. The performance curve of our proposed TGSR method under different hyper-parameter κ, i.e., the salient facial region number. (a) shows the experimental results of Exp.8(H→C) and (b) shows the experimental results of Exp.4(H→N).

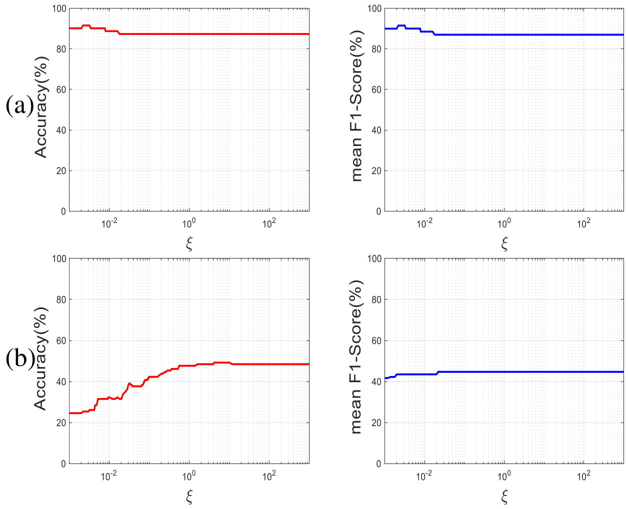

Fig. 4. The performance curve of our proposed TGSR method under different hyper-parameter ξ, i.e., weighting hyper-parameter of the MMD term. (a) shows the experimental results of Exp.1(H→V) and (b) shows the experimental results of Exp.12(N→C).

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[35]. 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.

3) Hyper-parameter Discussion: Two hyper-parameters are involved in solving the optimal regression matrix $\hat{\boldsymbol{C}}$ of proposed our 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 the model sensitiveness to hyper-parameters κ and ξ.

The number of salient facial regions. 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 our proposed 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. And it also verifies the effectiveness of salient facial region selection.

The weighting hyper-parameter of MMD term. 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 our proposed TGSR yield better performance. And the MMD term can effectively and stably improve model performance across a wide range of hyper-parameter ξ.

4) Visualization: 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 the AU definition of micro-expressions; 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 to select the salient facial regions to better cope with the Cross-Database Micro-Expression Recognition (CDMER) problem. In our proposed TGSR, salient facial region selection is achieved by the group feature selection from the raw face feature. This operation enables our proposed TGSR to 1) optimize the measurement and alleviate the difference between the source and target databases and 2) highlight the valid facial regions to enhance extracted features. In addition, this operation can also reduce computational costs while improving performance. Experiments and visualizations show that our proposed TGSR learns the discriminative facial regions 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] 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

[8] 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

[9] 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

[10] 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

[11] 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

[12] 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

[13] 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

[14] 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

[15] 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

[16] 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

[17] 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

[18] 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

[19] 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, 4

[20] 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

[21] 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. 2, 4

[22] Z. T. Qin and D. Goldfarb, "Structured sparsity via alternating direction methods." *Journal of Machine Learning Research*, vol. 13, no. 5, 2012. 3

[23] 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. 3

[24] 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. 5

[25] 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. 5

[26] 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. 5

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

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

[29] 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. 5

[30] 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. 5

[31] 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

[32] 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. 5

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

[34] 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

[35] 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