Sample Self-Revised Network for Cross-Dataset Facial Expression Recognition

Xiaolin Xu
School of Biological Science and Medical Engineering
Southeast University
Nanjing, China
xuxiaolin@seu.edu.cn

Wenming Zheng*
Key Laboratory of Child Development and Learning Science,
Ministry of Education
Southeast University
Nanjing, China
wenming_zheng@seu.edu.cn

Yuan Zong*
School of Biological Science and Medical Engineering
Southeast University
Nanjing, China
xhzongyuan@seu.edu.cn

Cheng Lu
School of Information Science and Engineering
Southeast University
Nanjing, China
cheng.lu@seu.edu.cn

Xingxun Jiang
School of Biological Science and Medical Engineering
Southeast University
Nanjing, China
jiangxingxun@seu.edu.cn

2022 International Joint Conference on Neural Networks (IJCNN) | 978-1-7281-8671-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/IJCNN55064.2022.9892500

***Abstract*—Facial images with low quality, subjective annotation, severe occlusion, and rare subject identity can lead to the existence of outlier samples in facial expression datasets. These outlier samples are usually far from the center of the dataset in the feature space, resulting in huge differences in feature distribution, which severely restricts the performance of cross-dataset facial expression recognition (FER). To eliminate the influence of outlier samples on cross-dataset FER, we propose an unsupervised domain adaptation (UDA) method called Sample Self-Revised Network (SSRN), which 1) dynamically detects the outlier level of each sample in the source domain to reduce the disturbance of outlier samples to the model training, as well as 2) adaptively revises outlier samples in the source domain to improve transferability of the learned features. Experimental results show that our SSRN outperforms both classic deep UDA methods and state-of-the-art cross-dataset FER results.**

***Index Terms*—Cross-dataset facial expression recognition, facial expression recognition, unsupervised domain adaptation, transfer learning**

I. Introduction

Facial expressions play an immeasurable role in interpersonal communication because they can visually convey human beings' emotional conditions. Therefore, increasing researchers have paid attention to the research of facial expression recognition (FER) and have proposed lots of effective methods [1]–[8]. Currently, most FER methods are designed and evaluated under an ideal assumption that the training and testing facial expression images come from the same dataset. However, in practical scenarios, the training and testing samples usually belong to different datasets. In this case, the performance of most aforementioned FER methods may drop sharply due to the feature distribution mismatch existing between the training and testing sets. This thus brings a greater challenge on FER tasks, i.e., cross-dataset FER.

So as to deal with cross-dataset FER, current mainstream approaches treat it as an Unsupervised domain adaptation (UDA) problem and design corresponding methods. For instance, Li et al. [9] proposed a Deep Emo-transfer Network (DETN), which utilizes Re-weighted Maximum Mean Discrepancy (MMD) [10] to improve the generalization performance of the model. Zhou et al. [11] proposed an Uncertainty-Aware cross-dataset facial Expression Transfer Network (UA-ETN), which can enhance the generalization ability of cross-dataset FER by aligning the marginal distribution and class-conditional distribution. Wang et al. [12] utilized additional data generated by the Generative Adversarial Network (GAN) to optimize the cross-dataset performance of FER. However, it should be pointed out that several interference factors shown in Fig. 1, e.g., subjective annotations, occlusion, illumination, and racial difference, may produce some outliers in the facial expression samples. These outlier samples inevitably break the modeling of the relationship between the source domain and its corresponding label information in dealing with cross-dataset FER, which leads to the UDA models failing to learn the discriminative facial expression features. In this case, the UDA models cannot cope with the cross-database FER tasks, although they successfully eliminate the feature distribution difference between the source and target domains with the help of a well-designed strategy. Therefore, it is important to consider the outlier samples in dealing with cross-dataset FER tasks.

Recently, outlier problems have been focused on conventional FER tasks, and numerous methods have been proposed to handle outlier samples. For example, Wang et al. [13] con-

*Corresponding authors

978-1-7281-8671-9/22/$31.00 ©2022 IEEE

sidered the outlier samples produced by the incorrect labeling and raised the Self-Cure Network (SCN) to cope with FER. The basic idea of SCN is to suppress these outlier samples in the model learning by seeking and relabeling them from the training facial expression samples. Zeng et al. [14] indicated the outlier samples caused by inconsistent annotations among different FER datasets and proposed an Inconsistent Pseudo Annotations to Latent Truth (IPA2LT) framework. IPA2LT artificially creates outlier samples by attaching multiple pseudo-labels to the samples, which are labeled by humans or the learned models, to learn the potential mapping between images and truth. By means of this latent mapping, the model can obtain better performance from multiple datasets with inconsistent annotations. Inspired by the success of the above works, in this paper, we take outlier samples in the source domain into consideration while exploring cross-dataset FER and further propose an effective network termed as Sample Self-Revised Network (SSRN). Specifically, the proposed SSRN consists of three essential modules: Outlier Perception Module, Outlier Perception Coefficient (OPC) Revision Module, and Feature Transfer Module. Given a batch of samples from source and target domains, respectively, the CNN Backbone can be used to extract facial features. Then the Outlier Perception and the OPC Revision Modules are used to dynamically perceive outlier samples in the source domain and mitigate the influence of outlier samples on the model training with the weighted cross-entropy loss [15]. After that, the well-designed Feature Transfer Module enforces the outlier samples to share the same or similar feature distribution with the source domain to adaptively revise the distribution of outlier samples in the source domain. In addition, features of source and target domains can also be aligned to learn more robust domain invariant features with the help of the Feature Transfer Module.

In summary, the main contributions in this paper include three folds:

1) To the best of our knowledge, this is the first work to consider the outlier samples in dealing with cross-dataset FER tasks, and we propose a novel deep domain adaptation method called Sample Self-Revised Network (SSRN), which fully considers the outliers existing in the source domains.

2) In the proposed SSRN model, we elaborately design a set of modules to respectively seek and revise outliers in source domains and align the revised source domain and the target one. Thus, the SSRN model can learn better domain-invariant features to describe facial expressions and have more promising performance in dealing with cross-dataset FER.

3) Extensive cross-dataset FER experiments across four public datasets are conducted to evaluate the proposed SSRN. Additional cross-dataset experiment beyond FER further indicates the strong applicability of our SSRN.

II. Proposed Method

To solve the influence of outlier samples on cross-dataset FER, we propose a simple yet effective method called Sample Self-Revise Network (SSRN). Firstly, the pipeline of the proposed SSRN is presented in subsection II-A, and then we introduce three important components of the SSRN in detail in subsections II-B–II-D below, respectively.

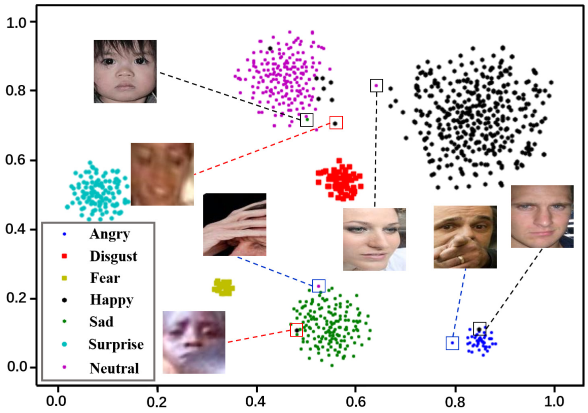

Fig. 1: Visualization of Outlier samples in RAF-DB dataset by t-SNE. We first randomly select 1000 samples from RAF-DB dataset and extract their 512-dimensional features in the last layer of ResNet-18. It is clear to see that several outlier samples indeed exist in the dataset, where the RED, BLUE, and BLACK Rectangle Marquees highlight the samples with low quality, subjective annotation, and severe occlusion, respectively.

A. Overview of SSRN

SSRN consists of three crucial modules: Outlier Perception, Outlier Perception Coefficient (OPC) Revision, and Feature Transfer shown in Fig.2. These modules aim to reduce or eliminate the influence of outlier samples on cross-dataset FER. Outlier Perception and OPC Revision Modules are utilized to dynamically detect outlier samples so as to suppress the contribution of outlier samples to the model training. Further, the Feature Transfer Module can revise outlier samples adaptively to obtain more robust domain invariant features. Detailed implementation will be described in the following part of this section. Before that, several necessary notations are introduced in the following, including source features and target features, which are denoted by $\mathbf{F}^s = \left[\mathbf{f}_1^s, \mathbf{f}_2^s, \cdots, \mathbf{f}_{N_s}^s\right] \in \mathbb{R}^{D \times N_s}$ and $\mathbf{F}^t = \left[\mathbf{f}_1^t, \mathbf{f}_2^t, \cdots, \mathbf{f}_{N_t}^t\right] \in \mathbb{R}^{D \times N_t}$. N_s and N_t are the number of source and target datasets, and D represents the dimension of feature vectors of source and target samples.

B. Outlier Perception

Outlier Perception is used to perceive the level of outlier samples in the source domain. It is generally recognized that facial images with low resolution or serious occlusion will be given a high level, and unambiguous images will be assigned to a low level. Outlier Perception contains a fully connected layer and a sigmoid activation function which takes the feature $\mathbf{F}^s$ as input and considers the output scalar $\alpha_i \, (\mathrm{i} \in 1, 2, \cdots, \mathrm{N})$ as the Outlier Perception Coeffi-

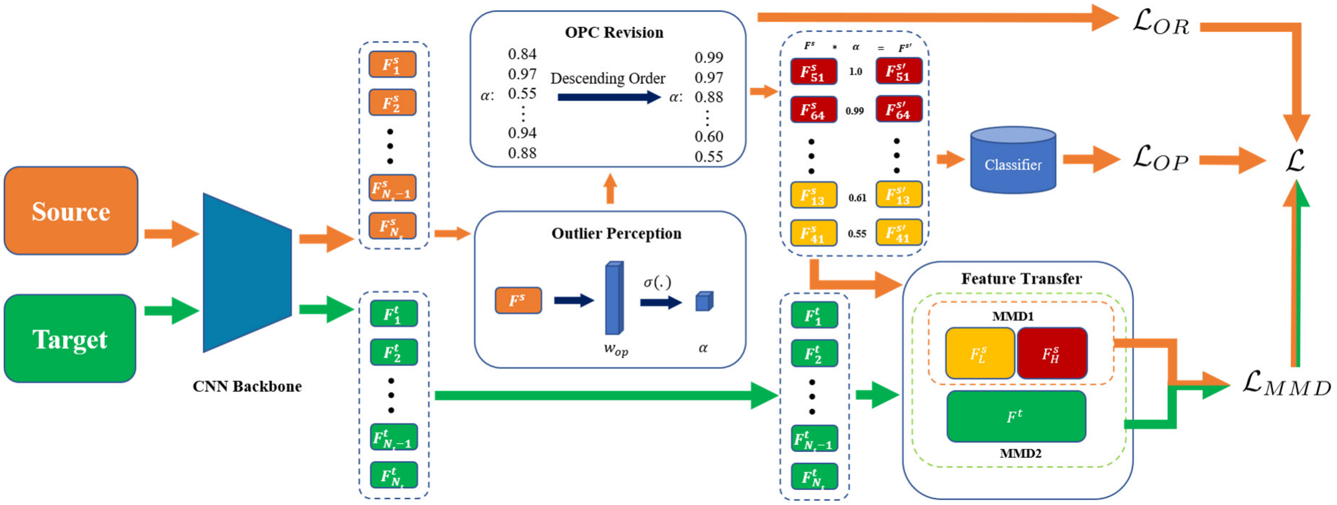

Fig. 2: The outline of SSRN: orange and green rectangular blocks represent the source and target domains. The red and yellow blocks indicate the features from high-coefficient group and low-coefficient group in the source domain. The orange and green flows represent processes for source and target domains, respectively.

cient (OPC) of the image. Its implementation is formulated as:

$$\alpha_i = \sigma\left(\mathbf{w}_{op}^{\mathrm{T}} \mathbf{f}_i^s\right), \tag{1}$$

where $\sigma(\cdot)$ represents the sigmoid activation function, and $\mathbf{w}_{op}$ represents parameters of the fully connected layer. The output α of the above formula is inversely proportional to the outlier level of each sample, which means the higher the image quality or the more consistent the annotation is, the higher α will be assigned; conversely, the lower the image quality is, the lower corresponding α will be.

A suitable loss function [15] is chosen to adjust the contribution of source data with different OPC. The aforementioned loss function is termed as Outlier Perception Loss (OP-Loss), which is formulated as,

$$q_{ij} = \frac{\mathrm{e}^{\alpha_i \mathbf{e}_j^{\mathrm{T}} \mathbf{W}_{cls}^{\mathrm{T}} \mathbf{f}_i^s}}{\sum_{k=1}^{C} \mathrm{e}^{\alpha_i \mathbf{e}_k^{\mathrm{T}} \mathbf{W}_{cls}^{\mathrm{T}} \mathbf{f}_i^s}}, \tag{2}$$

$$\mathcal{L}_{OP} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{C} p_{ij} \log q_{ij}, \tag{3}$$

where p_{ij} and q_{ij} represent the ground truth and the prediction of the j-th class of the i-th sample, respectively, $\mathbf{e}_j$ is a one-hot vector, whose j-th element is one and the others equal to zero. $\mathbf{W}_{cls}$ represents parameters of the classifier. By employing the weighted cross-entropy loss function, the model can reduce the interference of outlier samples to the model training and pay more attention to the samples with low outlier levels.

C. OPC Revision

The aforementioned Outlier Perception Module can perceive the level of outlier samples in the source domain. To further guarantee that the meaningful mapping relationship between facial expression features and outlier level can be learned correctly, the OPC Revision Module is proposed to revise and constrain the α generated from the Outlier Perception Module. The module first takes the OPC of each sample in the source domain as input and then arranges them in descending order. After that, it divides them into the high-coefficient and low-coefficient groups by a certain ratio β. In addition, a newly designed loss function, i.e., OPC Revision Loss (OR-Loss), is proposed to make sure that the mean value of the high-coefficient group is higher than the low-coefficient group with a certain margin, which is formulated as,

$$\mathcal{L}_{OR} = \max\left\{0, \mathit{Margin} - (\alpha_H - \alpha_L)\right\}, \tag{4}$$

where *Margin* represents the set margin between the mean value of the high-coefficient and low-coefficient groups. α_H and α_L represent mean values of the high-coefficient and low-coefficient groups calculated by

$$\alpha_H = \frac{1}{M} \sum_{i=1}^{M} \alpha_i \tag{5}$$

and

$$\alpha_L = \frac{1}{N_s - M} \sum_{i=M+1}^{N_s} \alpha_i, \tag{6}$$

where the number of samples in group high-coefficient and low-coefficient is $M = N_s \times \beta$ and $N_s - M$, respectively.

D. Feature Transfer

With the help of the Outlier Perception Module and OPC Revision Module, the model can learn the OPC of each sample relative to the sample feature center in the source domain. However, outlier samples could make the model learn the incorrect facial expression features and deter the model from aligning the features in the source and the target domain, which makes it more difficult for the model to learn domain invariant features. Therefore, the Feature Transfer Module is proposed to mitigate these problems. Before introducing the implementation of the Feature Transfer Module in detail, let us introduce some necessary theoretical bases.

1) Preliminary: As the most widely used loss function in domain adaptation, Maximum Mean Discrepancy (MMD) [10] is mainly applied for comparing distributions between two domains [16], which can be formulated as:

$$\mathrm{MMD}^2(X,Y)=\left\|\frac{1}{n}\sum_{i=1}^{n}\phi(x_i)-\frac{1}{m}\sum_{j=1}^{m}\phi(y_j)\right\|_{\mathcal{H}}^2, \quad (7)$$

where x_i and y_j represent the sample from domain X and Y, n and m represent the number of samples in the domain X and Y. $\mathcal{H}$ denotes the Reproducing Kernel Hilbert Space (RKHS). Projecting the data into $\mathcal{H}$ by the mapping function $\phi(\cdot)$, we can transform the inner product of function in the RKHS to the form of kernel function:

$$K(x,y)=\langle\phi(x),\phi(y)\rangle_{\mathcal{H}}, \quad (8)$$

where $K(x,y)$ represents the kernel function. In most UDA tasks [17], [18], the most commonly used kernel function is Gaussian kernel function:

$$K(x,y)=e^{\frac{-\|x-y\|^2}{2\sigma^2}}, \quad (9)$$

which can map the data to the infinite-dimensional space. [19] proposed the unbiased estimation expression of MMD after expanding the square of the (7):

$$\begin{aligned}\mathrm{MMD}^2(X,Y)=&\frac{1}{n^2}\sum_{i=1}^{n}\sum_{i'=1}^{n}K(x_i,x_{i'})\\&+\frac{1}{m^2}\sum_{j=1}^{m}\sum_{j'=1}^{m}K(y_j,y_{j'})\\&-\frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}K(x_i,y_j).\end{aligned} \quad (10)$$

In addition, we further replace MMD with single fixed kernel by multi-kernel MMD (MK-MMD) [10], which can be formulated as:

$$\mathcal{K}:=\left\{K=\sum_{u=1}^{d}\beta_u K_u:\beta_u\geq 0,\forall u\in\{1,\ldots,d\}\right\}. \quad (11)$$

MK-MMD obtains the optimal kernel by linearly weighting multiple kernels $\{k_u\}_{u=1}^{d}$ with weight β_u, which is more powerful in representation compared with the single-kernel MMD.

2) Feature Transfer Module: The Feature Transfer Module embeds MK-MMD in the task-specific feature layer of the deep network, which can be expressed as:

$$\mathcal{L}_{MMD}=\mathrm{MMD}\left(\mathbf{F}_H^s,\mathbf{F}_L^s\right)+\mathrm{MMD}\left(\mathbf{F}^s,\mathbf{F}^t\right), \quad (12)$$

where $\mathbf{F}_{\boldsymbol{H}}^{\boldsymbol{s}}$ and $\mathbf{F}_{\boldsymbol{L}}^{\boldsymbol{s}}$ represent the feature vectors of the high-coefficient and low-coefficient groups in the source domain. The first half of (12) reduces the feature distribution distance of high-coefficient group and low-coefficient group samples in the source domain, which revises outlier samples detected by Outlier Perception and OPC Revision Modules. The second half of the (12) further ensures that the target domain features can be aligned with the newly generated source domain features so that the model can learn more robust domain invariant features.

We accumulate (3), (4) and (12) as the final formulation of the proposed SSRN, which can be written as:

$$\mathcal{L}=\mathcal{L}_{OP}+\lambda\mathcal{L}_{MMD}+\gamma\mathcal{L}_{OR}, \quad (13)$$

where λ and γ are the trade-off factors to balance the proposed modules.

III. EXPERIMENT

In this section, we first describe four public facial expression datasets and then present our experiment protocol. Finally, we demonstrate the implementation details of our cross-dataset FER experiments.

A. Datasets

Four public available facial expression datasets, including FER2013 [20], RAF-DB [21], CK+ [22], and JAFFE [23] are applied to evaluate the proposed SSRN (shown in Fig.3).

FER2013 is a large-scale facial expression dataset consisting of 35,887 facial images of size 48×48 pixels. Each image is labeled by one of seven basic expression categories, i.e., Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral. The dataset is further divided into 28,709 training samples, 3,589 validation samples, and 3,589 testing samples. In the experiments, we adopt all the original facial images.

RAF-DB is a real-world facial expression dataset that contains 19,672 facial images collected from the Internet. We use its single-labeled subset consisting of 15,339 samples, which is divided into a training set of 12,271 images and a test set of 3,068 images, and each sample was assigned one of seven basic expressions.

Extended Cohn-Kanade (CK+) is a lab-controlled dataset that has 123 subjects and records their 593 facial expression video clips. Each video clip is annotated by one of six expressions, including Angry, Disgust, Fear, Happy, Sad, and Surprise. In the experiments, we extract the last peak frame from each labeled video clip to serve as expression samples and randomly choose the first frame from 50 sequences as the neutral ones [9].

JAFFE is also a lab-controlled dataset that contains 213 grayscale images from 10 Japanese female expressers. Expressers were asked to pose seven facial expressions. All images in the JAFFE are employed in our experiments.

Sample statistics of the aforementioned datasets are listed in Table I.

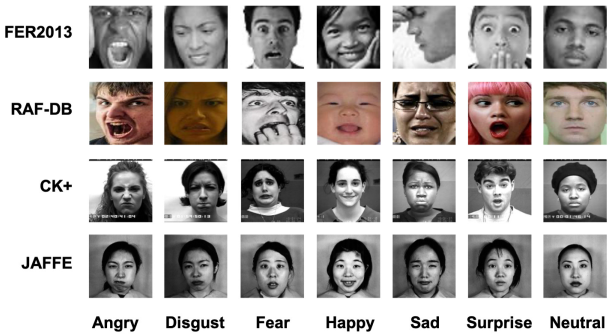

Fig. 3: Sample images from FER2013, RAF-DB, CK+ and JAFFE datasets.

TABLE I: Sample statistics of RAF-DB, JAFFE, CK+ and FER2013 for Experiments

Dataset	Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral	Total
FER2013	4953	547	5121	8989	6077	4002	6198	35887
RAF-DB	867	877	355	5957	2460	1619	3204	15339
CK+	45	59	25	69	28	83	50	359
JAFFE	30	29	32	31	31	30	30	213

B. Experiment Protocol

To evaluate the proposed SSRN, we resize all facial images to 112×112 pixels and then design a total of 12 experiments across the above four datasets in pairs, which are denoted by *R→J, R→C, R→F, J→R, J→C, J→F, C→R, C→J, C→F, F→R, F→J*, and *F→C*,where *R, J, C*, and *F* are the abbreviation of RAF-DB, JAFFE, CK+, and FER2013, and left and right sides of → represent the source domain and target domain respectively. As for the performance metric, we employ the recognition accuracy, which is calculated by T/N×100%, where T is the number of correct predictions and N is the total sample number in the target dataset.

C. Implementation Details

During the experiment, our SSRN is trained on 2 Nvidia Titan X GPUs, with the implementation in Pytorch. The CNN backbone of SSRN is ResNet-18 [24], which was pre-trained on ImageNet dataset [25] and the facial features with the dimension of 512 are extracted after the average pooling layer of ResNet-18. In the pre-processing stage, face images are detected and aligned by MTCNN [26], and further transformed to gray-scale. We argue the image by resizing the image to 128 × 128 and then randomly crop to 112 × 112, as well as adding horizontal flip with the probability of 50%. The division ratio β is set to 0.7, and the *Margin* represents the difference between the mean value of high and low groups is set to 0.15 by default. The ratio of $\mathcal{L}_{OP}$, $\mathcal{L}_{MMD}$ and $\mathcal{L}_{OR}$ will be discussed in the Evaluation of Trade-Off Parameters of section IV. Further, the influence of these three losses will be explored in the Ablation Study of section IV. We run our model using stochastic gradient descent (SGD) with an initial learning rate of 0.0001, which is divided by 10 after every 20 epochs, a momentum of 0.9, and a weight decay of 0.0005.

TABLE II: Results of cross-dataset FER experiments between RAF-DB, JAFFE, CK+ and FER2013.

Experiments	DANN [27]	DAN [18]	state-of-the-art	Ours
R→J	63.85	62.44	57.75 [9]	**67.61**
R→C	76.04	77.99	78.83 [9]	**82.45**
R→F	45.94	50.05	**52.37** [9]	50.83
J→R	36.52	**38.84**	-	**38.84**
J→C	44.85	68.25	65.01 [12]	**71.03**
J→F	**27.59**	25.14	-	25.05
C→R	39.18	36.75	-	**40.38**
C→J	31.92	51.17	51.64 [12]	**53.52**
C→F	29.26	28.52	-	**29.43**
F→R	59.81	**63.60**	-	62.08
F→J	52.11	54.46	50.70 [28]	**55.40**
F→C	59.33	63.23	-	**65.74**
Average	47.20	51.70	-	**53.53**

IV. Results and Discussion

Experimental results of the proposed SSRN are shown in this section. We first present the results of cross-dataset FER experiments between RAF-DB, JAFFE, CK+, and FER2013 in subsection IV-A. Then, we discuss the Trade-Off Parameters of three losses in subsection IV-B. After that, the influence of each module of SSRN is investigated in subsection IV-C. Finally, in subsection IV-D, we also apply our method to other UDA tasks beyond FER to illustrate the strong applicability of our method.

A. Results across RAF-DB, JAFFE, CK+, and FER2013

The results of 12 experiments across RAF-DB, JAFFE, CK+, and FER2013 datasets in pairs are shown in this subsection. To offer a fair comparison, we choose two well-performing deep domain adaption methods, i.e., DANN [27], and DAN [18] to conduct the experiments. We also compare our method with recent state-of-the-art methods and directly extract the results achieved by them under the same protocol from their corresponding literature. The experimental results of various UDA methods are presented in Table II. Several interesting findings can be observed from the experimental results.

Firstly, our SSRN achieved better results in 5 of 6 experiments compared with three state-of-the-art methods, which indicates that the proposed SSRN has a more powerful emotional discriminative ability in dealing with cross-dataset FER tasks.

Secondly, compared with the classical DAN and DANN methods, our SSRN outperforms the two traditional UDA methods in 9 out of 12 experiments and achieves the same results as DAN in *J→R* task. The average recognition accuracy of these three methods shows that SSRN is 6.33% higher than the DANN method and 1.83% higher than the classic DAN method, which more fairly and intuitively shows the superior performance of our SSRN in cross-dataset tasks. This is because, compared with traditional UDA methods, especially DAN, our method takes outlier samples in the source domain into consideration and proposes corresponding modules to alleviate its impact on cross-dataset FER tasks.

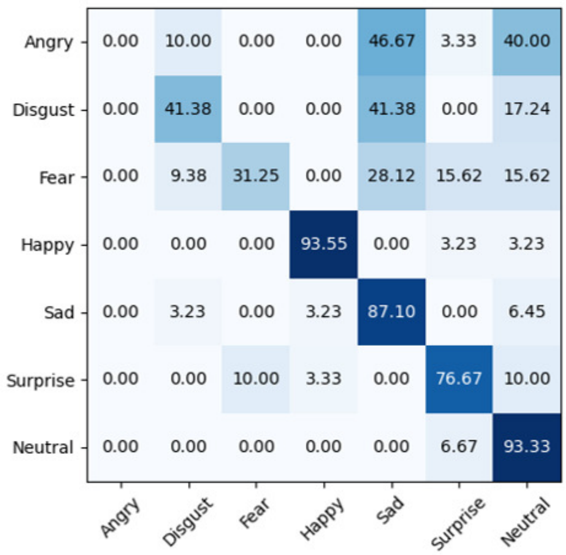

Fig. 4: The confusion matrices of baseline in the Experiment R→J.

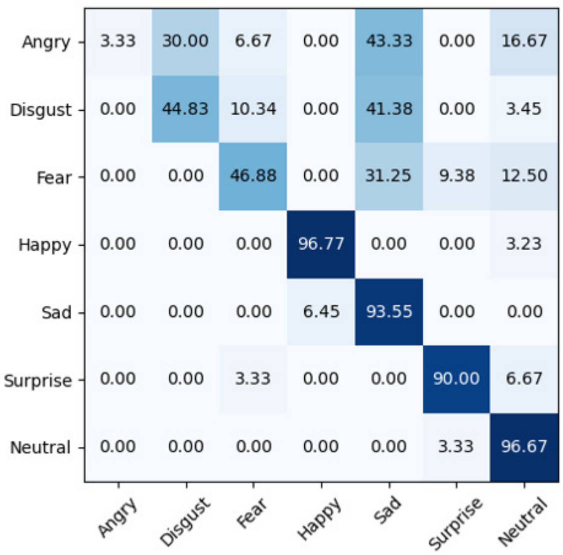

Fig. 5: The confusion matrices of SSRN in the Experiment R→J.

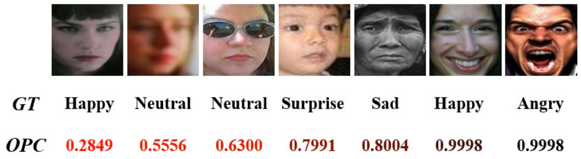
GT Happy Neutral Neutral Surprise Sad Happy Angry
OPC 0.2849 0.5556 0.6300 0.7991 0.8004 0.9998 0.9998

Fig. 6: Visualization of OPC in RAF-DB learned by SSRN: GT represents the ground truth of the corresponding samples.

Finally, In *R*→*F* and *F*→*R* experiments, our SSRN fails to achieve the best results, and the *R*-*F* experiment was 1.54% lower than DETN [9]. We speculate that this is due to DETN taking the serious class imbalance problem into consideration, which is ubiquitous in RAF-DB and FER2013 datasets. In addition, our SSRN performed poorly in experiments *J*→*F* and *J*→*R*, due to the small sample size and almost no outlier sample in the source domain, which hinders the performance of our SSRN.

Confusion matrices of the baseline method (ResNet-18) and SSRN on the cross-dataset FER experiment R→J are shown in Fig.4 and Fig.5 respectively. It is clear to see that our SSRN outperforms the baseline method in terms of all the facial expressions. More specifically, compared with the baseline method, the accuracies of Surprise and Fear achieved by SSRN have an increase of over 13%, which demonstrates that our SSRN can learn more discriminative features in dealing with cross-dataset FER.

In addition, we also visualize the OPC of samples in RAF-DB to investigate the effectiveness of our SSRN. The experimental results reported in Fig.6 show that our SSRN gives a lower OPC to the facial images with low quality, huge difference in personal attributes (e.g., age and ethnic), inconsistent annotation, and serious occlusion, while giving a higher OPC to facial images with clarity, no objection, and no occlusion. The experimental results indicate that the proposed SSRN can effectively suppress outlier samples in the source domain.

B. Evaluation of Trade-Off Parameters

We conduct the sensitivity analysis of trade-off parameters on the F→C experiment. In the first experiment, we investigate the impact of λ, which represents the trade-off between $\mathcal{L}_{OP}$ and $\mathcal{L}_{MMD}$. We set $\gamma = 0.1$ and $\lambda = [0 : 0.1 : 0.5]$ to conduct different cross-dataset FER experiments. The verification accuracies of these models are shown in the left side of Fig.7. It can be observed that the recognition accuracy of our SSRN varies slightly with respect to the change of the trade-off parameter λ. In the second experiment, we explore the impact of γ. We set $\lambda = 0.3$ and $\gamma = [0, 0.1, 0.3, 0.5, 0.8, 1.0]$ to learn different models. The verification accuracies of these models are shown in the right side of Fig.7. It can be observed that accuracies of our SSRN also remain largely stable across a wide range of the trade-off parameter γ, which demonstrated that our SSRN is less sensitive to the choice of trade-off parameters λ and γ.

C. Ablation Study

To better demonstrate the effect of each module of SSRN, an ablation study is conducted on *C*→*J* and *F*→*C*. Some conclusions can be observed from experimental results shown in Table III. First of all, compared with the baseline, which is composed of CNN backbone ResNet-18 in SSRN, a separate Outlier Perception Module has limited improvement on the performance of the model. This is because, without the constraint of the OPC Revision Module, the OPC generated by the Outlier Perception Module will become meaningless and become closer to one in value after many rounds of

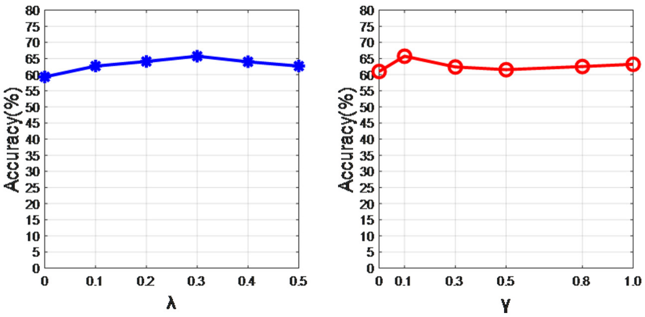

Fig. 7: Evaluation of hyper parameter λ and γ in the Experiment F→C.

TABLE III: Ablation Study in SSRN

Method	$C\!\to\!J(\%)$	$F\!\to\!C(\%)$
Baseline	47.42	57.66
$\mathcal{L}_{OP}$(SSRN)	47.42	59.61
$\mathcal{L}_{OP}+\mathcal{L}_{OR}$(SSRN)	49.77	61.28
$\mathcal{L}_{OP}+\mathcal{L}_{OR}+\mathcal{L}_{MMD}$(SSRN)	53.52	65.74

training. Secondly, adding the OPC Revision Module to the Outlier Perception Module can improve the performance, while the Feature Transfer Module can further enhance the performance of the model. Finally, the biggest improvement for one module is realized by adding the Feature Transfer Module, which Increases by 3.75% in $C\!\to\!J$ and 4.46% in $F\!\to\!C$. This indicates that the Feature Transfer Module plays the most important role in our SSRN.

D. Applicability of SSRN beyond FER

In addition to the above experiments, we are surprised that our method still has strong applicability to other UDA tasks. Additional experiments in other applications of domain adaptation have been added to indicate the strong applicability of our SSRN. Specifically, we conduct the experiments of domain adaptation handwritten digit recognition between the wildly used MNIST [29] and USPS [30]. To achieve this goal, we randomly select 10000 images from both datasets, and each category of the dataset has 1000 images to ensure the data balance of the samples. All the images are resized to 112×112 pixels. Note that we also create the outliers for both datasets by randomly selecting 10%, 20%, and 30% of the sample in the source domain and then adding random noise, performing Gaussian blur to images, and modifying labels. ResNet-18 (Baseline) [24] and classic UDA method DAN [18] are included in the comparison. Experimental results are shown in Table IV. It is clear to see that the performance of all the methods would decrease with respect to the increase of the proportion of outliers existing in the source domain. In addition, compared with the ResNet-18 and DAN, the proposed SSRN can promisingly improve the recognition accuracy in the target domain regardless of the proportion of outliers, which indicates that our method is robust to outliers in the source domain and also applicable to other UDA tasks besides cross-dataset FER.

TABLE IV: The evaluation of SSRN between USPS and MNIST. Outliers are synthesized and their proportion is presented in the first column.

Outlier	Experiment	ResNet-18 [24]	DAN [18]	SSRN
10%	U→M	75.2	83.39	**92.81**
	M→U	95.73	96.78	**98.39**
20%	U→M	73.1	79.13	**91.37**
	M→U	95.36	95.43	**97.65**
30%	U→M	72.31	73.77	**89.13**
	M→U	94.73	94.82	**97.31**

V. Conclusion

In this paper, a simple yet efficient method called SSRN has been proposed to solve the impact of outlier samples on model performance in cross-dataset FER tasks. With the help of the Outlier Perception and OPC Revision Modules, we are able to perceive the outlier level of the source data. Via providing different OPC to different samples in the source domain, the module can suppress the contribution of outlier samples to the model training and highlight the samples with a low outlier level. Furthermore, the Feature Transfer Module utilizes MK-MMD, on the one hand, constrains the distribution distance of high-coefficient group samples and low-coefficient group samples in the source domain to revise outlier samples in the source domain. On the other hand, it aligns features in the source and target domains to obtain more robust domain invariant features. Extensive experimental results indicate that our SSRN method achieves a better performance than some classic deep UDA methods and state-of-the-art cross-dataset FER methods, which demonstrate the effectiveness of SSRN.

Finally, we would like to discuss the shortcomings existing in our work, which are worth further investigating. First, our work only focuses on the influence of outlier samples in the source domain. In fact, outliers may also exist in the target domain. It would be better to simultaneously consider the outliers from both source and target domains in dealing with cross-dataset FER. Subsequently, the proposed SSRN seeks the outliers solely based on the label information of the source samples. However, in cross-dataset FER tasks, the label information of the target domain is not provided. Consequently, it is urgent to investigate other outlier perception methods such that the outliers from the target domain can be considered. In the future, we will further focus on the outlier samples in cross-dataset FER by considering the above two points.

VI. 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] Z. Yu and C. Zhang, "Image based static facial expression recognition with multiple deep network learning," in *Proceedings of the 2015 ACM on international conference on multimodal interaction*, 2015, pp. 435–442.

[2] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *2016 IEEE Winter conference on applications of computer vision (WACV)*. IEEE, 2016, pp. 1–10.

[3] X. Jia, X. Zheng, W. Li, C. Zhang, and Z. Li, "Facial emotion distribution learning by exploiting low-rank label correlations locally," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9841–9850.

[4] P. Ekman and W. V. Friesen, "Facial action coding system," *Environmental Psychology & Nonverbal Behavior*, 1978.

[5] W. Zheng, X. Zhou, C. Zou, and L. Zhao, "Facial expression recognition using kernel canonical correlation analysis (kcca)," *IEEE transactions on neural networks*, vol. 17, no. 1, pp. 233–238, 2006.

[6] R. Zhi, M. Flierl, Q. Ruan, and W. B. Kleijn, "Graph-preserving sparse nonnegative matrix factorization with application to facial expression recognition," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 41, no. 1, pp. 38–52, 2010.

[7] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, "Training deep networks for facial expression recognition with crowd-sourced label distribution," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 2016, pp. 279–283.

[8] A. Dhall, R. Goecke, S. Ghosh, J. Joshi, J. Hoey, and T. Gedeon, "From individual to group-level emotion recognition: Emotiw 5.0," in *Proceedings of the 19th ACM international conference on multimodal interaction*, 2017, pp. 524–528.

[9] S. Li and W. Deng, "Deep emotion transfer network for cross-database facial expression recognition," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 3092–3099.

[10] A. Gretton, D. Sejdinovic, H. Strathmann, S. Balakrishnan, M. Pontil, K. Fukumizu, and B. K. Sriperumbudur, "Optimal kernel choice for large-scale two-sample tests," in *Advances in neural information processing systems*. Citeseer, 2012, pp. 1205–1213.

[11] L. Zhou, X. Fan, Y. Ma, T. Tjahjadi, and Q. Ye, "Uncertainty-aware cross-dataset facial expression recognition via regularized conditional alignment," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 2964–2972.

[12] X. Wang, X. Wang, and Y. Ni, "Unsupervised domain adaptation for facial expression recognition using generative adversarial networks," *Computational intelligence and neuroscience*, vol. 2018, 2018.

[13] K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao, "Suppressing uncertainties for large-scale facial expression recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6897–6906.

[14] J. Zeng, S. Shan, and X. Chen, "Facial expression recognition with inconsistently annotated datasets," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 222–237.

[15] W. Hu, Y. Huang, F. Zhang, and R. Li, "Noise-tolerant paradigm for training face recognition cnns," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 887–11 896.

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

[17] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," *arXiv preprint arXiv:1412.3474*, 2014.

[18] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *International conference on machine learning*. PMLR, 2015, pp. 97–105.

[19] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 723–773, 2012.

[20] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee *et al.*, "Challenges in representation learning: A report on three machine learning contests," in *International conference on neural information processing*. Springer, 2013, pp. 117–124.

[21] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2852–2861.

[22] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *2010 ieee computer society conference on computer vision and pattern recognition-workshops*. IEEE, 2010, pp. 94–101.

[23] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with gabor wavelets," in *Proceedings Third IEEE international conference on automatic face and gesture recognition*. IEEE, 1998, pp. 200–205.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[25] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[26] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[27] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International conference on machine learning*. PMLR, 2015, pp. 1180–1189.

[28] G. Wen, Z. Hou, H. Li, D. Li, L. Jiang, and E. Xun, "Ensemble of deep neural networks with probability-based fusion for facial expression recognition," *Cognitive Computation*, vol. 9, no. 5, pp. 597–610, 2017.

[29] L. Deng, "The mnist database of handwritten digit images for machine learning research [best of the web]," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 141–142, 2012.

[30] J. J. Hull, "A database for handwritten text recognition research," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 16, no. 5, pp. 550–554, 1994.