

Article

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

Xiaolin Xu ^{1,2}, Yuan Zong ^{1,2,*}, Cheng Lu ¹ and Xingxun Jiang ^{1,2}

¹ Key Laboratory of Child Development and Learning Science of Ministry of Education, Southeast University, Nanjing 210096, China

² School of Biological Science and Medical Engineering, Southeast University, Nanjing 210096, China

* Correspondence: xhzyuan@seu.edu.cn

Abstract: Recently, cross-dataset facial expression recognition (FER) has obtained wide attention from researchers. Thanks to the emergence of large-scale facial expression datasets, cross-dataset FER has made great progress. Nevertheless, facial images in large-scale datasets 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 clustering center of the dataset in the feature space, thus resulting in considerable differences in feature distribution, which severely restricts the performance of most cross-dataset facial expression recognition methods. To eliminate the influence of outlier samples on cross-dataset FER, we propose the enhanced sample self-revised network (ESSRN) with a novel outlier-handling mechanism, whose aim is first to seek these outlier samples and then suppress them in dealing with cross-dataset FER. To evaluate the proposed ESSRN, we conduct extensive cross-dataset experiments across RAF-DB, JAFFE, CK+, and FER2013 datasets. Experimental results demonstrate that the proposed outlier-handling mechanism can reduce the negative impact of outlier samples on cross-dataset FER effectively and our ESSRN outperforms classic deep unsupervised domain adaptation (UDA) methods and the recent state-of-the-art cross-dataset FER results.

Keywords: cross-dataset facial expression recognition; facial expression recognition; unsupervised domain adaptation; transfer learning



Citation: Xu, X.; Zong, Y.; Lu, C.; Jiang, X. Enhanced Sample Self-Revised Network for Cross-Dataset Facial Expression Recognition. *Entropy* **2022**, *24*, 1475. <https://doi.org/10.3390/e24101475>

Academic Editors: Andrea Prati, Luis Javier García Villalba and Vincent A. Cicirello

Received: 7 August 2022

Accepted: 10 October 2022

Published: 17 October 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Facial expressions reveal human emotional conditions and mental outlook visually. Therefore, automatic facial expression recognition (FER) has great application prospects in the fields of human–computer interface (HCI) [1] and medical care [2]. Over the last decade, increasing researchers have paid attention to the research of facial expression recognition. Several effective methods [3–6] have driven the progress of FER. At the same time, the emergence of a variety of facial expression datasets [7–11] has also greatly improved FER performance.

Although FER has made great strides in recent years, it should be pointed out that most of the FER methods mentioned above are designed and evaluated under an ideal assumption that the training and testing of facial expression images come from the same dataset. Under this assumption, the training and testing samples follow the same or similar distribution. However, in practical scenarios, the training and testing samples usually belong to different datasets, which means the training and testing samples may have a difference in the pose, illumination, background, and level of expressiveness [12]. 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, thus bringing researchers a greater challenge (i.e., cross-dataset FER) compared with the traditional FER.

However, due to the influence of the annotation environment and cultural background, the annotators of the FER dataset may produce extreme subjective annotation deviation [12].

In addition, several other interference factors, such as sample quality (e.g., occlusion and sample clarity) and racial bias [13], may also produce some outliers in facial expression datasets as shown in Figure 1. Wang et al. [14] declared that outlier samples in facial expression datasets could be harmful to the model for learning useful facial expression features. Xu et al. [15] also revealed the performance degradation of other cross-dataset tasks, through extensive experiments, caused by learning incorrect outlier samples. Similarly, in dealing with cross-dataset FER, these outlier samples can also inevitably break the modeling of the relationship between the source domain and its corresponding label information. Therefore, the unsupervised domain adaptation (UDA) models could fail to learn the discriminative facial expression features. In this case, it is hard for the UDA models to cope with the cross-database FER tasks, although they successfully eliminate the feature distribution difference between the source and target domains with well-designed strategies. Therefore, it is crucial to consider the outlier samples in dealing with cross-dataset FER tasks.

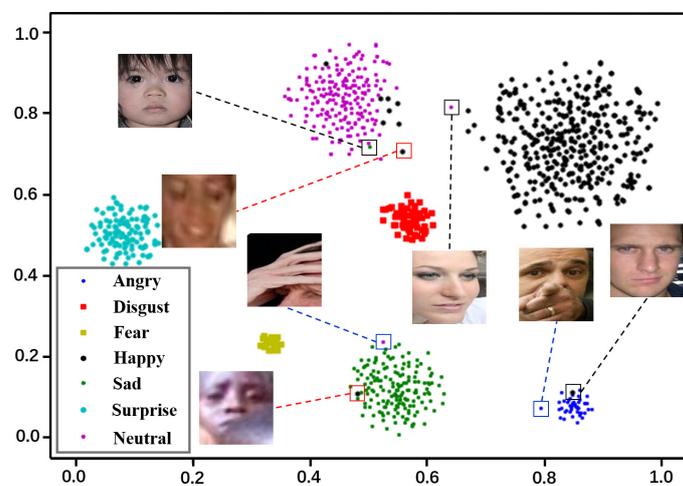


Figure 1. Visualization of outlier samples in RAF-DB dataset by t-SNE. We randomly select 1000 samples from the RAF-DB dataset and extract their 512-dimensional features in the last layer of ResNet-18. It is clear to see that several outlier samples indeed exist in the dataset, where the RED, BLUE, and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

To deal with this problem, we proposed an effective network termed as sample self-revised network (SSRN) in our conference work of [15]. SSRN is committed to learning a feature extractor that is more robust to both source and target domains. By reducing the influence of outlier samples in the source domain on model training and aligning the target domain with the revised source domain at the feature level, SSRN can obtain the domain-invariant features for both source and target domains. Precisely, 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 these outliers on the model training. After that, the well-designed feature transfer module enforces the outlier samples to share the same feature distribution with other samples in the source domain so as to revise these outlier samples. In addition, features of source and target domains can also be aligned to learn more robust domain invariant features with the feature transfer module.

In this paper, we strengthen our conference work on SSRN [15], and further propose the enhanced sample self-revised network (ESSRN). Specifically, we modify our original

SSRN from two aspects. First, the class imbalance of the FER dataset can sharply degrade the domain adaptation performance [16]. Inspired by [17], we add conditional maximum mean discrepancy (MMD) [18] while dealing with outliers revision and domain adaptation in the feature transfer module. In addition to aligning the marginal distribution considered in previous works, we further align the class-conditional distribution by narrowing the feature distance of each category between the source and the target domains. Second, we added a feature reconstruction layer in the revision process. The feature reconstruction layer performs linear reconstruction of outlier samples at the feature level so that the revised samples can mitigate the inconsistency of modeling the relationship between the source domain and their label information. In summary, besides the original contributions in our preliminary work [15], in this paper, we further improve the outliers revision function and the unsupervised domain adaptation process to learn better domain-invariant features to describe facial expressions. Therefore, we achieve more promising performance in dealing with cross-dataset FER.

The rest of this paper is organized as follows. We review recent works about cross-dataset facial expression recognition and methods for outliers in Section 2. In Section 3, we introduce our proposed ESSRN in detail. Section 4 presents four public facial expression datasets and our experiment protocol. Extensive experiments and evaluations are provided in Section 5. Finally, the work of this paper is concluded in Section 6.

2. Related Works

2.1. Cross-Dataset Facial Expression Recognition

To deal with cross-dataset FER, current mainstream approaches treat it as an unsupervised domain adaptation (UDA) problem and design corresponding methods [12,16,19–25]. For instance, Yan et al. [19] proposed an unsupervised domain adaptive dictionary learning method (UDADL). Combining dictionary learning with unsupervised domain adaptation, UDADL achieved satisfactory running time and recognition accuracy results. Zheng et al. [20] proposed a transductive transfer subspace learning method (TTRLSR). Combining a labeled image set from the source domain with an unlabeled auxiliary image set from the target domain, TTRLSR jointly learns a discriminative subspace to deal with cross-dataset FER. In the work of [21,22], Zong et al. aimed to learn a domain regenerator that regenerates source samples (target samples) sharing the same or similar feature distribution as the target samples (source samples). Other approaches based on deep learning have also achieved promising results on cross-dataset FER. Wang et al. [23] utilized additional data generated by the generative adversarial network (GAN) to optimize the cross-dataset performance of FER. Zhou et al. [24] proposed an uncertainty-aware cross-dataset facial expression transfer network (UA-ETN). Aligning the marginal and class-conditional distribution, UA-ETN can enhance the generalization ability of cross-dataset FER. In the work of [16], Li et al. revealed the influence of class imbalance on cross-dataset FER and proposed a re-weighting method termed as deep emo-transfer network (DETN), which utilizes re-weighted maximum mean discrepancy (MMD) [18] to improve the generalization performance of the model. Based on the above work, Li et al. explored the bias in FER datasets in depth in the work of [12]. They pointed out the influence of inconsistent conditional probability distributions in source and target domains on cross-dataset FER. To this end, they proposed emotion-conditional adaptation network (ECAN), which considers both marginal and conditional distributions in cross-dataset FER. Xie et al. [25] proposed an adversarial graph representation adaptation (AGRA) that combines graph representation propagation with adversarial learning to learn local and global joint features to realize cross-dataset FER better.

2.2. Methods for Outliers

Many solutions have been proposed to deal with outlier samples [14,26–28] existing in all kinds of datasets. Among these datasets, the noisy samples generated by subjective annotation are the most widely studied. Aiming at the problem of inconsistent annotation in dataset annotation, Dawid et al. [26] used the EM algorithm to estimate the quality of

each annotator in the crowdsourcing process to determine the labels. This study provided a theoretical basis for the later calibration of database labels. Li et al. [28] proposed a unified distillation framework, which uses information from a small clean dataset and label relations in a knowledge graph, to avoid learning from noisy labels. In contrast, other approaches refuse to employ a clean set of data. Ref. [29] assumed additional constraints or distributions on the noise labels to learn the relationship between the noisy samples and the corresponding latent labels.

Recently, outlier problems have been focused on conventional FER tasks, and numerous methods have been proposed to handle outlier samples. For example, as the first work to consider outlier samples in FER, Zeng et al. [27] indicated the outlier samples were 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 labeled by humans or the learned models to learn the potential mapping between images and truth. Employing this latent mapping, the model can obtain better performance from multiple datasets with inconsistent annotations. Recently, Wang et al. [14] considered 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.

3. Methods

Outliers can seriously damage the extraction of domain invariant features on cross-dataset FER. To address these issues, we propose a simple yet effective method called enhanced sample self-revised network (ESSRN). The description of ESSRN in this section is arranged as follows. Firstly, the overview of the proposed ESSRN is presented in Section 3.1. Then we review the previous work of SSRN [15]. Three important components of SSRN are introduced in detail in Sections 3.2–3.4, respectively. Based on SSRN, we are able to propose the ESSRN in this paper. Therefore, after that, we demonstrate the improvements of ESSRN over SSRN in Section 3.5. At last, the specific optimization process of ESSRN is given in Section 3.6.

3.1. Overview of ESSRN

In this part, we introduce the overview of ESSRN. ESSRN consists of three crucial modules: outlier perception, outlier perception coefficient (OPC) revision, and feature transfer, which are shown in Figure 2. These modules work together to implement the proposed novel outliers handling mechanism, whose aim is to seek outlier samples and then suppress them in dealing with cross-dataset FER. Firstly, the outlier perception module seeks outlier samples dynamically and suppresses the model from learning incorrect knowledge from outlier samples. Secondly, the OPC revision module includes a constraint to strengthen the outliers detection ability of the outlier perception module. Finally, the feature transfer module can obtain more robust domain invariant features in dealing with cross-dataset FER by aligning the marginal and class-conditional distribution of the revised source and target domains. 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 $\mathbf{F}^s = [\mathbf{f}_1^s, \mathbf{f}_2^s, \dots, \mathbf{f}_{N_s}^s] \in \mathbb{R}^{D \times N_s}$ and $\mathbf{F}^t = [\mathbf{f}_1^t, \mathbf{f}_2^t, \dots, \mathbf{f}_{N_t}^t] \in \mathbb{R}^{D \times N_t}$. 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.

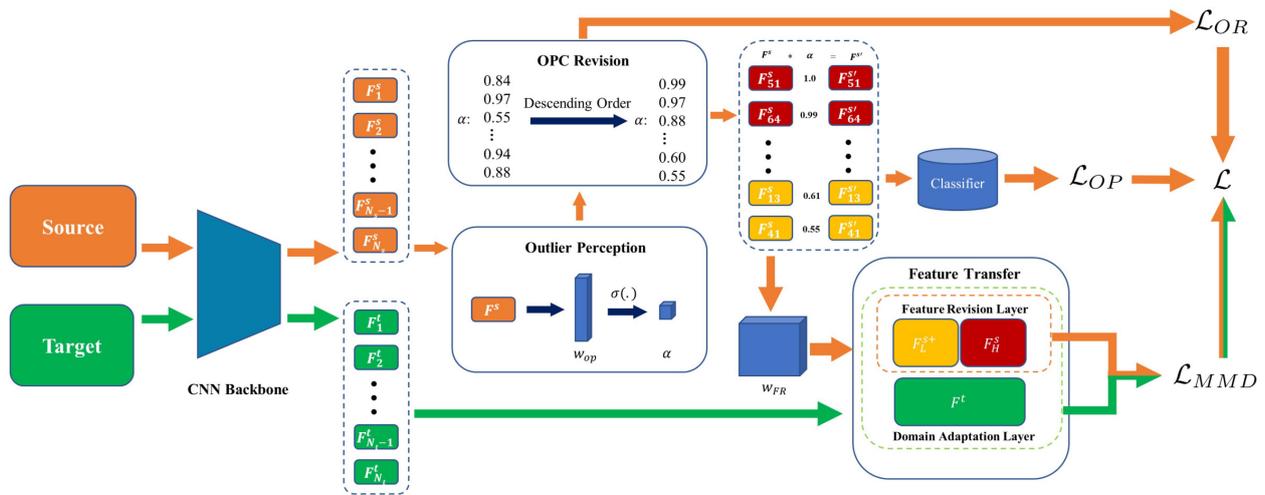


Figure 2. The illustration of ESSRN: orange and green rectangular blocks represent the source and target domains. The red and yellow blocks indicate the features from high-coefficient group and low-coefficient group in the source domain. The orange and green flows represent processes for source and target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

3.2. Outlier Perception

Outlier perception perceives the level of outlier samples in the source domain and allows the SSRN to apply attention to different data during the model training adaptatively. 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. To realize the outlier perception, Outlier Perception employs a fully connected layer and a sigmoid activation function which takes the feature F^s as input and considers the output scalar $\alpha_i (i \in 1, 2, \dots, N)$ as the outlier perception coefficient (OPC) of each image. Its implementation is formulated as

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

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.

In order to allow the model to learn dynamically from the samples at different weights, a suitable loss function [30], i.e., outlier perception loss, is chosen to adjust the contribution of source data with different OPC. Outlier perception loss (OP-Loss) can be formulated as

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

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

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 are 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.

3.3. 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 , among which the samples in the low-coefficient group can be regarded as the outlier samples we mentioned above. In addition, a newly designed loss function, i.e., OPC revision loss (OR-Loss), is proposed to ensure that the low-coefficient and high-coefficient groups can be separated. The formulation follows can guarantee that the mean value of the high-coefficient group is higher than the low-coefficient group with a certain margin,

$$\mathcal{L}_{OR} = \max\{0, Margin - (\alpha_H - \alpha_L)\}, \quad (4)$$

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

$$\alpha_H = \frac{1}{M} \sum_{i=1}^M \alpha_i \quad (5)$$

and

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

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

3.4. 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 correctly. 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 target domains. 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.

3.4.1. Preliminary

As the most widely used loss function in domain adaptation, maximum mean discrepancy (MMD) [18] is mainly applied for comparing distributions between two domains [31], which can be formulated as

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

where x_i and y_j represent the sample from the 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, the most commonly used kernel function is the Gaussian kernel function:

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

which can map the data to the infinite-dimensional space. The unbiased estimation expression of MMD after expanding the square of the (7) can be formulated as follows:

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

In addition, we further replace MMD with a single fixed kernel by multi-kernel MMD (MK-MMD) [18], which can be formulated as follows:

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

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.

3.4.2. 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 follows:

$$\mathcal{L}_{MMD} = \text{MMD}(\mathbf{F}_H^s, \mathbf{F}_L^s) + \text{MMD}(\mathbf{F}^s, \mathbf{F}^t), \tag{12}$$

where \mathbf{F}_H^s and \mathbf{F}_L^s represent the feature vectors of the high-coefficient and low-coefficient groups in the source domain. The loss function of the feature transfer module is composed of two parts. 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 the outlier perception module. The second half of (12) further ensures that the target domain features can be aligned with the revised source domain features so that the model can learn more robust domain invariant features.

3.5. From SSRN to ESSRN

Through the above content, we recalled the specific implementation details of SSRN. In this article, we improved SSRN based on previous work and further propose ESSRN. To be more specific, in the implementation of ESSRN, we not only paid attention to the marginal distribution difference between the source and target domains, but also minimized the conditional distribution difference between the corresponding categories of the source and target domains. We retained the structure of outlier perception and OPC revision modules in the original SSRN, and mainly improved the feature transfer module. In the feature transfer module of ESSRN, we newly proposed two functional structures: feature revision layer and domain adaptation layer. With the help of the feature transfer module, the feature transfer module could first better revise the outlier samples in the source domain at the feature level so that the revised source domain could keep a relatively consistent distribution. Then, the domain adaptation layer further narrowed the distribu-

tion difference between the revised source and target domains to extract robust domain invariant features.

3.5.1. Feature Revision Layer

In order to further reduce the influence of outlier samples on domain invariant feature learning, we add an additional feature revision layer in the feature transfer module compared with the previous work. The feature revision layer first employs a fully connected layer to revise the outliers in the source domain linearly. Then, we propose a loss function based on MK-MMD to force the generated feature to share the same or similar feature distribution with the source domain. The proposed loss function jointly minimizes the MMD distance of marginal and conditional probability distributions between outliers and the source domain to better revise the outliers at the feature level. The entire feature revision layer can be formulated as follows:

$$\mathbf{F}_L^{s+} = \mathbf{w}_{FR}^T \mathbf{F}_L^s, \tag{13}$$

$$\begin{aligned} \mathcal{L}_{MMD1} = & \left\| \frac{1}{M} \sum_{i=1}^M \phi(f_H^{s_i}) - \frac{1}{N_s - M} \sum_{i=1}^{N_s - M} \phi(f_L^{s_i+}) \right\|_{\mathcal{H}}^2 \\ & + \frac{1}{C} \sum_{j=1}^C \left\| \frac{1}{M_j} \sum_{i=1}^{M_j} \phi(f_H^{s_{ij}}) - \frac{1}{N_{s_j} - M_j} \sum_{i=1}^{N_{s_j} - M_j} \phi(f_L^{s_{ij}+}) \right\|_{\mathcal{H}}^2, \end{aligned} \tag{14}$$

where \mathbf{F}_L^{s+} represents the revised outlier feature, and \mathbf{w}_{FR} is the parameters of the fully connected layer for feature revision. $f_H^{s_{ij}}$ and $f_L^{s_{ij}+}$ denote the i th feature with the j th class from \mathbf{F}_H^s and \mathbf{F}_L^{s+} respectively. M_j and N_{s_j} are the corresponding sample numbers satisfying $\sum_{j=1}^C M_j = M$ and $\sum_{j=1}^C N_{s_j} = N_s$.

3.5.2. Domain Adaptation Layer

In the previous work of [15], we embedded MK-MMD in the task-specific feature layer of the feature transfer module, which only aligns the source and target domains' marginal distribution and ignores the distribution difference between source and target domains in each category. Inspired by the success of the work of [17], we propose a more adaptable loss function in the domain adaptation layer. Thus, we can simultaneously narrow the difference between the marginal and class-conditional distribution between source and target domains. The newly proposed loss function can be expressed as

$$\begin{aligned} \mathcal{L}_{MMD2} = & \left\| \frac{1}{N_s} \sum_{i=1}^{N_s} \phi(f_i^s) - \frac{1}{N_t} \sum_{i=1}^{N_t} \phi(f_i^t) \right\|_{\mathcal{H}}^2 \\ & + \frac{1}{C} \sum_{j=1}^C \left\| \frac{1}{N_{s_j}} \sum_{i=1}^{N_{s_j}} \phi(f_{ij}^s) - \frac{1}{N_{t_j}} \sum_{i=1}^{N_{t_j}} \phi(f_{ij}^t) \right\|_{\mathcal{H}}^2, \end{aligned} \tag{15}$$

where f_{ij}^s and f_{ij}^t denote the i th feature with the j th class from \mathbf{F}^s and \mathbf{F}^t , respectively. N_{s_j} and N_{t_j} are the corresponding sample numbers satisfying $\sum_{j=1}^C N_{s_j} = N_s$ and $\sum_{j=1}^C N_{t_j} = N_t$.

We accumulate (14) and (15) as the loss function of the feature transfer module, i.e., MMD loss, which can be formulated as follows:

$$\mathcal{L}_{MMD} = \mathcal{L}_{MMD1} + \mathcal{L}_{MMD2}. \tag{16}$$

The final formulation of the proposed ESSRN is accumulated by (3), (4) and (7), which can be written as

$$\mathcal{L} = \mathcal{L}_{OP} + \lambda \mathcal{L}_{MMD} + \gamma \mathcal{L}_{OR}, \tag{17}$$

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

3.6. Optimization of ESSRN

In this section, how the network is optimized is explained in detail. In order to calculate the \mathcal{L}_{MMD} , we need to give the pseudo label of the target domain. Therefore, we use the alternated direction method to optimize the whole ESSRN.

Specifically, the model optimization process is presented in Algorithm 1. We first randomly initialize the parameters of ESSRN, i.e., w_{op} , w_{cls} and w_{FR} , and then predict the pseudo labels L_p^t of the target domain. Subsequently, we perform the following two major steps until the total loss \mathcal{L} is equal to zero or the number of iterations reaches the maximal iteration N_{iter} :

1. Calculating the total loss \mathcal{L} according to L_p^t , and updating the network parameters θ by the optimization algorithm, e.g., SGD and Adam.
2. Fix parameters θ , and update the pseudo target domain labels L_p^t .

Algorithm 1 The detailed procedures for optimization of ESSRN

Input: Source Image Features: $\mathbf{F}^s = [\mathbf{f}_1^s, \mathbf{f}_2^s, \dots, \mathbf{f}_{N_s}^s]$,
 Target Image Features: $\mathbf{F}^t = [\mathbf{f}_1^t, \mathbf{f}_2^t, \dots, \mathbf{f}_{N_t}^t]$,
 Trade-off Parameters: λ and γ ,
 Learning Rate: lr ,
 Maximal Iterations: N_{iter}

Output: Optimal Network Parameters: $\theta = \{w_{op}, w_{cls}, w_{FR}\}$ and Target Labels: L_p^t

- 1: Initialize the network parameters: θ , total loss: $\mathcal{L} = \mathcal{L}_{OP} + \lambda\mathcal{L}_{MMD} + \gamma\mathcal{L}_{OR}$, and iteration indicator: $iter = 0$
- 2: **while** $\mathcal{L} \neq 0$ and $iter < N_{iter}$ **do**
- 3: $iter = iter + 1$;
- 4: Fix the network parameters: θ , Predict the target domain pseudo label L_p^t ;
- 5: Fix the target domain pseudo label L_p^t , calculate \mathcal{L} ;
- 6: Update the network parameters: θ ;
- 7: $\nabla_{\theta} \leftarrow \frac{\partial(\mathcal{L}_{OP} + \lambda\mathcal{L}_{MMD} + \gamma\mathcal{L}_{OR})}{\partial\theta}$;
- 8: $\theta^{n+1} \leftarrow \theta^n - lr * \nabla_{\theta}$;
- 9: **end while**

4. Experiment

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

4.1. Datasets

Four public available facial expression datasets, including FER2013 [9], RAF-DB [10], CK+ [8], and JAFFE [7], are applied to evaluate the proposed ESSRN.

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 18,709 training samples, 3589 validation samples, and 3589 testing samples. In the experiments, we adopted 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 3068 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 seven

expressions, including angry, disgust, fear, happy, sad, contempt, and surprise. In the experiments, to be relatively consistent with other datasets, we only used emotional data for six of these categories, i.e., angry, disgust, fear, happy, sad, and surprise, as well as neutral. We extracted the last three peak frames from the video clips labeled with one of six target categories to serve as expression samples and chose the first frame from these labeled sequences as the neutral expression.

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

The aforementioned datasets have significant differences in pose, illumination, background, level of expressiveness, and image quality, which can be seen in the typical dataset images in Figure 3. Detailed sample statistics of these datasets are listed in Table 1.



Figure 3. Typical sample images under seven categories from FER2013, RAF-DB, CK+ and JAFFE datasets presented.

Table 1. 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	35,887
RAF-DB	867	877	355	5957	2460	1619	3204	15,339
CK+	135	177	75	207	84	249	309	1236
JAFFE	30	29	32	31	31	30	30	213

4.2. Experiment Protocol

To evaluate the proposed ESSRN, we resized all facial images to 112×112 pixels and then designed a total of 12 experiments across the above four datasets in pairs, which are denoted by $R \rightarrow J$, $R \rightarrow C$, $R \rightarrow F$, $J \rightarrow R$, $J \rightarrow C$, $J \rightarrow F$, $C \rightarrow R$, $C \rightarrow J$, $C \rightarrow F$, $F \rightarrow R$, $F \rightarrow J$, and $F \rightarrow C$, where R , J , C , and F are the abbreviations of RAF-DB, JAFFE, CK+, and FER2013, and the left and right sides of \rightarrow represent the source domain and target domain, respectively. As for the performance metric, we employed the recognition accuracy, which is calculated by $T/N \times 100\%$, where T is the number of correct predictions and N is the total sample number in the target dataset. In addition, all feature extraction networks in the experiment are with the widely used ResNet-18 [32].

4.3. Implementation Details

In this part, we illustrate our experiment implementation in detail, including pre-processing, model training, and hyper-parameters setting stages. In the pre-processing stage, face images are detected and aligned by MTCNN [33] and further transformed to

gray-scale. We adjusted the image by resizing the image to 128×128 and then randomly cropping to 112×112 , as well as adding the horizontal flip with the probability of 50%. For the training details, our ESSRN is trained on one Tesla A10 GPU, with the implementation in Pytorch.

The CNN backbone ResNet-18 [32] was pre-trained on ImageNet dataset [34] and the facial features with the dimension of 512 are extracted after the average pooling layer of ResNet-18. We ran our model using stochastic gradient descent (SGD) with an initial learning rate of 0.0001 for the backbone parameters and an initial learning rate of 0.01 for other parts, which are divided by 10 after every 20 epochs, a momentum of 0.9, and a weight decay of 0.0005. For the hyper-parameters setting, the division ratio β was set to 0.7, and the *margin* represents the difference between the mean value of high and low groups was 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 5. Further, the influence of these three losses will be explored in the ablation study of Section 5.

5. Results and Discussion

In this section, we present the experimental results of the proposed ESSRN. We first show the accuracies of our ESSRN across RAF-DB, JAFFE, CK+, and FER2013 in Section 5.1. Then, the sensitivity analysis of the trade-off parameters is discussed in Section 5.2. Finally, the influence of three modules of ESSRN is investigated in Section 5.3.

5.1. 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 relatively fair comparison, we also chose two well-performing deep domain adaption methods, i.e., DANN [35] and DAN [36], to conduct the experiments. In addition, ResNet-18 was chosen as the backbone network for all the comparable experiments to provide a fairer comparison. At last, we also compared our method with recent state-of-the-art cross-dataset FER methods. The experimental results of various UDA methods are presented in Table 2. Several interesting findings can be observed from the experimental results.

Table 2. Results of cross-dataset FER experiments between RAF-DB, JAFFE, CK+ and FER2013. The method with the highest accuracy in each experiment is indicated in bold.

Experiments	DANN [35]	DAN [36]	DETN [16]	ICID [37]	ECAN [12]	AGRA [38]	SSRN [15]	ESSRN
$R \rightarrow J$	63.85	62.44	52.11	48.83	52.11	61.03	67.61	63.85
$R \rightarrow C$	74.76	75.49	64.19	67.44	66.51	77.52	80.99	80.83
$R \rightarrow F$	45.94	50.05	42.01	53.00	50.76	54.94	50.83	50.98
$J \rightarrow R$	36.52	38.84	26.77	16.27	27.13	39.71	38.84	40.11
$J \rightarrow C$	48.46	65.94	31.01	24.03	40.31	82.17	63.03	68.85
$J \rightarrow F$	27.59	25.14	27.03	18.54	27.06	28.44	25.05	28.54
$C \rightarrow R$	36.97	43.68	23.26	19.73	26.04	35.44	45.67	44.52
$C \rightarrow J$	49.30	56.81	16.54	27.23	24.24	49.30	56.81	57.75
$C \rightarrow F$	31.14	33.08	18.74	23.77	19.33	30.89	29.98	34.54
$F \rightarrow R$	59.81	63.60	57.14	60.71	59.35	65.12	62.08	66.59
$F \rightarrow J$	52.11	54.46	33.65	38.50	37.86	48.16	55.40	65.26
$F \rightarrow C$	71.52	76.21	72.09	68.22	65.89	80.95	69.82	79.53
Average	49.83	53.81	38.71	38.86	41.38	54.47	53.84	56.78

Firstly, compared with the classical DAN and DANN methods, our ESSRN outperforms two traditional UDA methods in all experiments. The average recognition accuracy of these three methods shows that ESSRN is 6.95% higher than the DANN method and 2.97% higher than the classic DAN method, which shows the superior performance of our ESSRN in cross-dataset tasks more fairly and intuitively. 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.

Secondly, our ESSRN achieved better results in 9 of 12 experiments compared with the recent state-of-the-art cross-dataset FER methods (i.e., DETN [16], ICID [37], ECAN [12], and AGRA [38]), indicating the effectiveness of the proposed outliers handling mechanism in dealing with cross-dataset FER tasks.

Thirdly, compared with the previous SSRN, the proposed ESSRN achieved better results in 9 of the 12 experiments and improved by 2.94% in average accuracy. These results indicated that the newly added feature revision layer and the domain adaptation layer can improve the cross-dataset FER ability by paying extra attention to the difference in class-conditional distribution.

Fourthly, we believe that the source dataset largely determined the final performance in cross-dataset facial expression recognition tasks. Therefore, we calculated the average performance of the 12 experiments above with the same source domains, as shown in Table 3. We compared DAN, which generally has good performance in the classical UDA method, with our ESSRN method and calculated the performance increase based on DAN. According to the experimental results, when the source domains are in-the-wild datasets (e.g., RAF-DB and FER2013), the experimental performance is much higher than that of the lab-controlled datasets. This is because these in-the-wild datasets have considerable advantages in the sample size and diversity, making the model learn robust information. To our surprise, compared with DAN, our ESSRN also made greater progress on the experiments whose source domain is RAF-DB or FER2013. One possible reason for this phenomenon is that, compared with the lab-controlled datasets, the datasets in the wild are prone to generate outlier samples caused by subjective annotation, image occlusion, and other uncontrollable reasons. However, our model is quite sensitive to outlier samples in the training set and is good at reducing or even eliminating the impact of outlier samples on the cross-dataset FER.

Finally, compared with AGRA [38], one of the most currently available comprehensive and state-of-the-art methods, our ESSRN fails to achieve the best results in the $R \rightarrow F$, $J \rightarrow C$, and $F \rightarrow C$ experiments. We speculated that this is due to AGRA paying extra attention to the local information of the face based on the landmarks, and the common facial features can provide more detailed information for cross-dataset facial expression recognition. In addition, among these experiments, our ESSRN performed poorly in the $J \rightarrow C$ experiment, which we believe is due to the small sample size and almost no outlier sample in the source domain, which hinders the performance of our ESSRN.

In addition, we also visualize the OPC of samples in RAF-DB to investigate the effectiveness of our ESSRN. The experimental results reported in Figure 4 show that our ESSRN gives a lower OPC to facial images with low quality, huge differences in personal attributes (e.g., age and ethnicity), 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 ESSRN can effectively perceive outlier samples in the training set and reduce the proportion of these outlier samples in the model training.

Table 3. Average performance of experiments with the same source domains. The third column and the fourth column respectively represent the average accuracy of DAN and our ESSRN with the same source domain, and the last column indicates the increment of our ESSRN compared to DAN.

Source Set	Target Set	DAN [36]	ESSRN	Increment
FER2013	Average	64.76	70.46	5.70
RAF-DB	Average	62.66	65.22	2.56
JAFFE	Average	43.31	45.83	2.52
CK+	Average	44.52	45.60	1.08



GT	Happy	Neutral	Neutral	Disgust	Neutral	Neutral	Angry	Angry
OPC	0.0468	0.1656	0.2461	0.4016	0.5401	0.7431	0.9170	0.9937

Figure 4. Visualization of OPC in RAF-DB learned by ESSRN: GT represents the ground truth of the corresponding samples.

5.2. Evaluation of Trade-Off Parameters

In this part, we conduct the sensitivity analysis of trade-off parameters on the experiment, which treats RAF-DB as the source domain and CK+ as the target domain. 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 $\gamma = 0.6$ and $\lambda = [0.0001, 0.001, 0.01, 0.1, 1]$ to conduct different cross-dataset FER experiments. The verification accuracies of these models are shown on the left side of Figure 5. It can be observed that properly choosing the value of the trade-off parameter λ can improve the recognition accuracy. In the second experiment, we explore the impact of γ , which indicates the trade-off between \mathcal{L}_{OP} and \mathcal{L}_{OR} . We set $\lambda = 0.01$ and $\gamma = [0, 0.2, 0.4, 0.6, 0.8, 1.0]$ to learn different models. The verification accuracies of these models are shown in the right side of Figure 5. It can be observed that the accuracies of our ESSRN remain largely stable across a wide range of the trade-off parameter γ , which demonstrates that our ESSRN is less sensitive to the choice of trade-off parameters λ and γ .

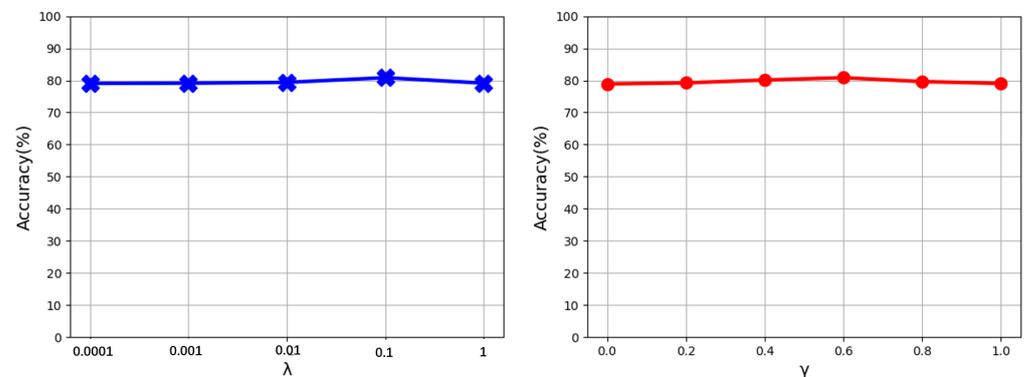


Figure 5. Evaluation of hyper parameter λ and γ on the experiment which treats RAF-DB as the source domain and CK+ as the target domain.

5.3. Ablation Study

An ablation study is conducted to demonstrate the effect of each module of our ESSRN. Four datasets are selected as the source domain, respectively, and one of the datasets is randomly selected as the target domain to conduct the ablation experiments. Some conclusions can be observed from experimental results shown in Table 4. First of all, compared to the baseline in the first row, which ResNet-18 conducts, each module in ESSRN improves the performance of cross-dataset expression recognition. Then, the second row in Table 4 reveals that ESSRN has limited improvement when the model only has the outlier perception module. This is because the outlier perception module can work properly only with the constraint of the OPC revision module. Otherwise, the OPC generated by the outlier perception module will become meaningless and become closer to one in value after many rounds of training. Thirdly, the added OPC revision module improves the performance of ESSRN, which is consistent with the previous analysis, while the feature transfer module can further enhance the model's performance. Finally, it can be easily

found that the biggest improvements are all achieved by the feature transfer module in four experiments, which increases by 2.82% in $R \rightarrow J$, 8.74% in $J \rightarrow C$, 4.23% in $C \rightarrow J$ and 3.04% in $F \rightarrow R$, respectively. This indicates the great advantages of applying our well-designed UDA module.

Table 4. Ablation study in ESSRN.

Method	$R \rightarrow J$ (%)	$J \rightarrow C$ (%)	$C \rightarrow J$ (%)	$F \rightarrow R$ (%)
Baseline	59.15	45.79	49.01	62.60
\mathcal{L}_{OP} (ESSRN)	60.09	52.02	50.23	62.78
$\mathcal{L}_{OP} + \mathcal{L}_{OR}$ (ESSRN)	61.03	60.11	53.52	63.55
$\mathcal{L}_{OP} + \mathcal{L}_{OR} + \mathcal{L}_{MMD}$ (ESSRN)	63.85	68.85	57.75	66.59

6. Conclusions

In this paper, we proposed the enhanced sample self-revised network (ESSRN) with novel outliers handling mechanism to solve the impact of outlier samples in cross-dataset FER tasks, which hinders the UDA models from learning the discriminative facial expression features. Our ESSRN can perceive the outlier level of the source data with the help of the outlier perception and OPC revision modules. Via tagging different OPC to different samples in the source domain, the proposed ESSRN is able to weaken or strengthen the contribution of outliers or other samples to the model training, respectively. Furthermore, the feature transfer module first 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 at the feature level. Then, the feature transfer module aligns the marginal and class-conditional distribution of the source and target domains with obtaining more robust domain invariant features. Extensive cross-dataset FER experiments between CK+, JAFFE, RAF-DB and FER2013 datasets were conducted to evaluate the performance of our ESSRN. Experimental results indicated that the proposed outliers handling mechanism can reduce the negative impact of outlier samples on cross-dataset FER effectively and our ESSRN achieves a better performance than the traditional deep UDA methods and most state-of-the-art cross-dataset FER methods.

Author Contributions: Conceptualization, X.X. and Y.Z.; methodology, X.X. and Y.Z.; software, X.X., C.L. and X.J.; validation, X.X. and Y.Z.; formal analysis, X.X. and Y.Z.; investigation, X.X., Y.Z. and C.L.; resources, Y.Z.; data curation, X.X.; writing—original draft preparation, X.X.; writing—review and editing, Y.Z.; visualization, X.X.; supervision, Y.Z.; project administration, X.X.; funding acquisition, Y.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by the Natural National Science Foundation of China (NSFC) under the Grants U2003207, 61902064, and 62076195, in part by the Jiangsu Frontier Technology Basic Research Project under the Grant BK20192004, and in part by the Zhishan Young Scholarship of Southeast University.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Cowie, R.; Douglas-Cowie, E.; Tsapatsoulis, N.; Votsis, G.; Kollias, S.; Fellenz, W.; Taylor, J.G. Emotion recognition in human-computer interaction. *IEEE Signal Process. Mag.* **2001**, *18*, 32–80. [[CrossRef](#)]
2. Edwards, J.; Jackson, H.J.; Pattison, P.E. Emotion recognition via facial expression and affective prosody in schizophrenia: A methodological review. *Clin. Psychol. Rev.* **2002**, *22*, 789–832. [[CrossRef](#)]
3. Li, Y.; Zeng, J.; Shan, S.; Chen, X. Occlusion aware facial expression recognition using CNN with attention mechanism. *IEEE Trans. Image Process.* **2018**, *28*, 2439–2450. [[CrossRef](#)] [[PubMed](#)]
4. Li, S.; Deng, W. Deep facial expression recognition: A survey. *IEEE Trans. Affect. Comput.* **2020**, *13*, 1195–1215. [[CrossRef](#)]

5. Ruan, D.; Yan, Y.; Lai, S.; Chai, Z.; Shen, C.; Wang, H. Feature decomposition and reconstruction learning for effective facial expression recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 21–25 June 2021; pp. 7660–7669.
6. Ge, H.; Zhu, Z.; Dai, Y.; Wang, B.; Wu, X. Facial expression recognition based on deep learning. *Comput. Methods Programs Biomed.* **2022**, *215*, 106621. [[CrossRef](#)] [[PubMed](#)]
7. Lyons, M.; Akamatsu, S.; Kamachi, M.; Gyoba, J. Coding facial expressions with gabor wavelets. In Proceedings of the 3rd IEEE International Conference on Automatic Face and Gesture Recognition, Nara, Japan, 14–16 April 1998; IEEE: Piscataway, NJ, USA, 1998; pp. 200–205.
8. Lucey, P.; Cohn, J.F.; Kanade, T.; Saragih, J.; Ambadar, Z.; Matthews, I. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, San Francisco, CA, USA, 13–18 June 2010; IEEE: Piscataway, NJ, USA, 2010; pp. 94–101.
9. Goodfellow, I.J.; Erhan, D.; Carrier, P.L.; Courville, A.; Mirza, M.; Hamner, B.; Cukierski, W.; Tang, Y.; Thaler, D.; Lee, D.H.; et al. Challenges in representation learning: A report on three machine learning contests. In Proceedings of the International Conference on Neural Information Processing, Daegu, Korea, 3–7 November 2013; Springer: Berlin/Heidelberg, Germany, 2013; pp. 117–124.
10. Li, S.; Deng, W.; Du, J. 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, Abu Dhabi, United Arab Emirates, 25–28 October 2020; IEEE: Piscataway, NJ, USA, 2017; pp. 2852–2861.
11. Kossaiji, J.; Walecki, R.; Panagakis, Y.; Shen, J.; Schmitt, M.; Ringeval, F.; Han, J.; Pandit, V.; Toisoul, A.; Schuller, B.; et al. Sewa db: A rich database for audio-visual emotion and sentiment research in the wild. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 1022–1040. [[CrossRef](#)] [[PubMed](#)]
12. Li, S.; Deng, W. A deeper look at facial expression dataset bias. *IEEE Trans. Affect. Comput.* **2020**, *13*, 881–893. [[CrossRef](#)]
13. Wang, M.; Deng, W.; Hu, J.; Tao, X.; Huang, Y. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 692–702.
14. Wang, K.; Peng, X.; Yang, J.; Lu, S.; Qiao, Y. Suppressing uncertainties for large-scale facial expression recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 6897–6906.
15. Xu, X.; Zheng, W.; Zong, Y.; Lu, C.; Jiang, X. Sample Self-Revised Network for Cross-Dataset Facial Expression Recognition. In Proceedings of the International Joint Conference on Neural Networks, Padua, Italy, 18–23 July 2022; pp. 63–76.
16. Li, S.; Deng, W. Deep emotion transfer network for cross-database facial expression recognition. In Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR), Union, NJ, USA, 15–17 August 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 3092–3099.
17. Long, M.; Wang, J.; Ding, G.; Sun, J.; Yu, P.S. Transfer feature learning with joint distribution adaptation. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; IEEE: Piscataway, NJ, USA, 2013; pp. 2200–2207.
18. Gretton, A.; Sejdinovic, D.; Strathmann, H.; Balakrishnan, S.; Pontil, M.; Fukumizu, K.; Sriperumbudur, B.K. Optimal kernel choice for large-scale two-sample tests. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; Citeseer: Princeton, NJ, USA, 2012; pp. 1205–1213.
19. Yan, K.; Zheng, W.; Cui, Z.; Zong, Y. Cross-database facial expression recognition via unsupervised domain adaptive dictionary learning. In Proceedings of the International Conference on Neural Information Processing, Kyoto, Japan, 16–21 October 2016; Springer: Cham, Switzerland, 2016; pp. 427–434.
20. Zheng, W.; Zong, Y.; Zhou, X.; Xin, M. Cross-domain color facial expression recognition using transductive transfer subspace learning. *IEEE Trans. Affect. Comput.* **2016**, *9*, 21–37. [[CrossRef](#)]
21. Zong, Y.; Huang, X.; Zheng, W.; Cui, Z.; Zhao, G. Learning a target sample re-generator for cross-database micro-expression recognition. In Proceedings of the 25th ACM International Conference on Multimedia, Mountain View, CA, USA, 23–27 October 2017; pp. 872–880.
22. Zong, Y.; Zheng, W.; Huang, X.; Shi, J.; Cui, Z.; Zhao, G. Domain regeneration for cross-database micro-expression recognition. *IEEE Trans. Image Process.* **2018**, *27*, 2484–2498. [[CrossRef](#)] [[PubMed](#)]
23. Wang, X.; Wang, X.; Ni, Y. Unsupervised domain adaptation for facial expression recognition using generative adversarial networks. *Comput. Intell. Neurosci.* **2018**, *2018*, 7208794. [[CrossRef](#)] [[PubMed](#)]
24. Zhou, L.; Fan, X.; Ma, Y.; Tjahjadi, T.; Ye, Q. Uncertainty-aware Cross-dataset Facial Expression Recognition via Regularized Conditional Alignment. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 2964–2972.
25. Xie, Y.; Chen, T.; Pu, T.; Wu, H.; Lin, L. Adversarial graph representation adaptation for cross-domain facial expression recognition. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 1255–1264.
26. Dawid, A.P.; Skene, A.M. Maximum likelihood estimation of observer error-rates using the EM algorithm. *J. R. Stat. Soc. Ser. (Appl. Stat.)* **1979**, *28*, 20–28. [[CrossRef](#)]
27. Zeng, J.; Shan, S.; Chen, X. Facial expression recognition with inconsistently annotated datasets. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 222–237.

28. Li, Y.; Yang, J.; Song, Y.; Cao, L.; Luo, J.; Li, L.J. Learning from noisy labels with distillation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1910–1918.
29. Jiang, L.; Zhou, Z.; Leung, T.; Li, L.; Fei-Fei, L. Mentornet: Regularizing very deep neural networks on corrupted labels. *arXiv* **2017**, arXiv:1712.05055.
30. Hu, W.; Huang, Y.; Zhang, F.; Li, R. Noise-tolerant paradigm for training face recognition CNNs. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 11887–11896.
31. Borgwardt, K.M.; Gretton, A.; Rasch, M.J.; Kriegel, H.P.; Schölkopf, B.; Smola, A.J. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics* **2006**, *22*, e49–e57. [[CrossRef](#)] [[PubMed](#)]
32. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
33. Zhang, K.; Zhang, Z.; Li, Z.; Qiao, Y. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process. Lett.* **2016**, *23*, 1499–1503. [[CrossRef](#)]
34. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Li, F.-F. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; IEEE: Piscataway, NJ, USA, 2009; pp. 248–255.
35. Ganin, Y.; Lempitsky, V. Unsupervised domain adaptation by backpropagation. In Proceedings of the International Conference on Machine Learning, PMLR, Lille, France, 7–9 July 2015; pp. 1180–1189.
36. Long, M.; Cao, Y.; Wang, J.; Jordan, M. Learning transferable features with deep adaptation networks. In Proceedings of the International Conference on Machine Learning, PMLR, Lille, France, 7–9 July 2015; pp. 97–105.
37. Ji, Y.; Hu, Y.; Yang, Y.; Shen, F.; Shen, H.T. Cross-domain facial expression recognition via an intra-category common feature and inter-category distinction feature fusion network. *Neurocomputing* **2019**, *333*, 231–239. [[CrossRef](#)]
38. Chen, T.; Pu, T.; Wu, H.; Xie, Y.; Liu, L.; Lin, L. *Cross-Domain Facial Expression Recognition: A Unified Evaluation Benchmark and Adversarial Graph Learning*; IEEE Transactions on Pattern Analysis and Machine Intelligence: Piscataway, NJ, USA, 2021; p. 1. [[CrossRef](#)]