electronics

MDPI

Article

Multi-Representation Joint Dynamic Domain Adaptation Network for Cross-Database Facial Expression Recognition

Jingjie Yan [1], Yuebo Yue [1], Kai Yu [1], Xiaoyang Zhou [2,3,*], Ying Liu [4], Jinsheng Wei [1] and Yuan Yang [5]

[1] Jiangsu Key Laboratory of Intelligent Information Processing and Communication Technology, College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China; yanjingjie@njupt.edu.cn (J.Y.); 1022010402@njupt.edu.cn (Y.Y.); 1221014130@njupt.edu.cn (K.Y.); 20220182@njupt.edu.cn (J.W.)
[2] School of Information Science and Engineering, Southeast University, Nanjing 210096, China
[3] China Mobile Zijin (Jiangsu) Innovation Research Institute Co., Ltd., Nanjing 211189, China
[4] China Mobile Communications Group Jiangsu Co., Ltd., Nanjing Branch, Nanjing 211135, China; liuyingnj2@js.chinamobile.com
[5] School of Instrument Science and Engineering, Southeast University, Nanjing 210096, China; yangyuan@seu.edu.cn
* Correspondence: zxyjobs51@163.com

Abstract: In order to obtain more fine-grained information from multiple sub-feature spaces for domain adaptation, this paper proposes a novel multi-representation joint dynamic domain adaptation network (MJDDAN) and applies it to achieve cross-database facial expression recognition. The MJDDAN uses a hybrid structure to extract multi-representation features and maps the original facial expression features into multiple sub-feature spaces, aligning the expression features of the source domain and target domain in multiple sub-feature spaces from different angles to extract features more comprehensively. Moreover, the MJDDAN proposes the Joint Dynamic Maximum Mean Difference (JD-MMD) model to reduce the difference in feature distribution between different subdomains by simultaneously minimizing the maximum mean difference and local maximum mean difference in each substructure. Three databases, including eNTERFACE, FABO, and RAVDESS, are used to design a large number of cross-database transfer learning facial expression recognition experiments. The accuracy of emotion recognition experiments with eNTERFACE, FABO, and RAVDESS as target domains reach 53.64%, 43.66%, and 35.87%, respectively. Compared to the best comparison method chosen in this article, the accuracy rates were improved by 1.79%, 0.85%, and 1.02%, respectively.

Keywords: transfer learning; facial expression recognition; multi-representation joint dynamic domain adaptation network; Joint Dynamic Maximum Mean Difference

Citation: Yan, J.; Yue, Y.; Yu, K.; Zhou, X.; Liu, Y.; Wei, J.; Yang, Y. Multi-Representation Joint Dynamic Domain Adaptation Network for Cross-Database Facial Expression Recognition. *Electronics* **2024**, *13*, 1470. https://doi.org/10.3390/electronics13081470

Academic Editor: Seong G. Kong

Received: 18 March 2024
Revised: 10 April 2024
Accepted: 11 April 2024
Published: 12 April 2024

1. Introduction

As one of the research directions in the field of human–computer interaction, emotion recognition has attracted much attention. Emotion is a key factor in understanding human behavior. In emotion recognition, facial expression recognition develops a major role in the domain of human–computer interaction and computer vision. With the progress of computer hardware and the vigorous development of the era of big data, human emotion poses a new challenge to computers and human–computer interaction. Although many research advances have been made in facial expression and emotion recognition, there is still room to improve the accuracy and efficiency of facial emotion judgments. This will open up many new opportunities for the technology-driven future society, especially in the fields of artificial intelligence-driven medical care, robotics, security monitoring, and corporate marketing, which have broad application prospects [1–4]. The process of facial expression recognition usually involves the use of computer technology to extract face information from images. Firstly, the face detection algorithm is used to recognize

and locate the face in the image. Subsequently, the image is preprocessed to improve its quality and prepare it for further analysis. Then, through the feature extraction method, important facial features are identified, and finally, emotion classification is carried out. Facial expression recognition can aid healthcare institutions in identifying individuals' mental stress issues, thereby facilitating their ability to cope with the challenges pertaining to their psychological well-being. Moreover, this technology can also offer support to businesses, sales, and other service industries by analyzing consumers' facial expressions for the development of corresponding marketing strategies [5–12].

When conducting experiments, existing facial emotion datasets have some limitations, including insufficient data volume, low image quality, and insufficient diversity of expression labels. In addition, accessing some datasets has become more difficult due to copyright and image right restrictions. Even if an individual wants to build a dataset, it takes a lot of time, effort, and cost to take photos and annotate information [13–16]. Some researchers in this field have carried out related research to solve these problems. In Section 2, we present relevant research on cross-database facial expression recognition.

Cross-database expression recognition can reduce the difference in sample distribution between different databases, but conducting cross-database expression experiments requires hardware support and a large amount of computing resources. At the same time, it also faces technical challenges such as inconsistent labels between different databases [14,17]. In the field of deep learning, cross-database transfer learning has emerged as a prominent research direction, with numerous effective methods being developed. Drawing inspiration from the Multi-Representation Adaptation Network (MRAN) approach [18], this paper proposes a hybrid structure composed of multiple sub-structure feature extraction units for multi-representation feature extraction. It maps the original facial expression images into multiple distinct sub-feature spaces, enabling the alignment of image features from different perspectives in both source and target domains. Previous methods in each sub-feature space solely focus on aligning the global distributions of data between source and target domains while neglecting the unique details within each category that are present in both domains. To extract more fine-grained information from expressions, this paper proposes a multi-representation joint dynamic domain adaptation network (MJDDAN) that performs joint dynamic domain adaptation on source and target domain data within each sub-feature space for cross-database facial expression recognition research.

The main contributions of this paper are as follows:

(1) A novel Joint Dynamic Maximum Mean Difference (JD-MMD) model integrating a subdomain and global domain is proposed, not only enabling adaptation between different domains but also the incorporation of fine-grained information within the same class.
(2) The acquisition of image features from multiple angles enables a more comprehensive description of the features. By incorporating these features into the joint dynamic domain adaptive model, we propose a multi-representation joint dynamic domain adaptive network (MJDDAN) that effectively addresses feature distribution discrepancies across different domains, thereby enhancing model performance and adaptability.
(3) The experimental results demonstrate that the domain adaptive method based on MJDDAN exhibits significant advantages in the task of facial expression recognition. This is validated through experiments conducted on the RAVDESS [19], FABO [20], and eNTERFACE [21] datasets. In comparison to other methods, MJDDAN achieves superior implementation outcomes.

The subsequent sections are organized as follows: The Section 2 provides an overview of the relevant studies conducted in the field of cross-database identification research. The Section 3 presents the multi-representation joint dynamic domain adaptive network (MJDDAN) and outlines the network's overall structure and principles of each module. The Section 4 provides a detailed description of the experiments, including information on the experimental environment, dataset selection, experimental details, procedures, and

evaluation metrics. Subsequently, an analysis and demonstration of the experimental results is presented to validate the effectiveness of our proposed method through comparisons with other approaches. Finally, in the Section 5, we summarize our proposed methodology by reviewing its core ideas and objectives while summarizing the key steps and techniques employed. Furthermore, we highlight the advantages and innovations offered by our approach.

2. Related Work

The TMMLDL migration model, proposed by Ni and his colleagues [22], integrates the techniques of dictionary learning and metric learning. This study also investigates the utilization of pairwise constraints and global structural information to ensure effective transfer learning across diverse domains. The proposed method by Zhang et al. [23], known as Joint Local-Global Discriminative Subspace Transfer Learning (LGDSTL), employs joint local-global graphs to quantify the dissimilarity between different domains, with the objective of achieving a balanced knowledge transfer from global graphs in each database and local discriminative geometric structures. Transfer Component Analysis (TCA) [24] is one of the classical transfer learning methods. It is a method based on subspace projection, which maps the data of the source domain and target domain into a shared feature space to realize knowledge transfer between the domains. It has achieved good results in many practical applications, which proves its practicability and effectiveness in solving the problem of domain adaptation. GFK (Geodesic Flow Kernel) [25] is a transfer learning method utilizing manifold learning. The GFK method uses the manifold distance between the covariance matrices of the source and target domains to construct a new kernel that can be used for classification. Specifically, the GFK method learns a new kernel function by finding the minimum manifold distance between the source domain and the target domain, which can be classified over the target domain. Subspace Alignment (SA) [26] is a transfer learning method utilizing kernel normalization, which aims to minimize the difference between the source domain and target domain through subspace alignment, thereby improving the performance of transfer learning.

In addition to traditional transfer learning methods, many transfer learning methods based on deep learning have also been proposed in recent years. A DANN (Domain Adversarial Neural Network) [27] uses the idea of domain classifier and adversarial training to achieve feature transfer and self-adaptation among different domains by minimizing the difference between the source domain and target domain. This method can help the feature representation learned by the model to have strong generalization performance to achieve better performance in the target domain. A Deep Subdomain Adaption Network (DSAN) [28] proposes an idea of self-adaptation between local subdomains to improve the effect of transfer learning. This approach recognizes that the differences between the source domain and the target domain are not global but exist in local subdomains. Therefore, a DSAN introduces the local maximum mean difference (LMMD) to measure the degree of difference between local subdomains. Ren et al. [29] developed a representational learning model that is transferable and can improve recognition performance. The method models bionic face representations as approximately stable and structured representations to identify commonalities between source and target domains. This method significantly improves the recognition performance by improving the grouping effect of the feature output and introducing a reasonable method to emphasize and share the discriminant scale and direction. Long et al. [30] verify that when inputting data with different attributes, they can complement each other and explore their role in making the learning model robust to domain differences. As a result, they propose a universal framework called Graph Co-Regularized Transfer Learning (GTL), which integrates various matrix factorization models. Specifically, the main objective of GTL is to extract latent common factors by preserving the statistical properties of different domains and further optimizing them using the geometric structure of each domain to mitigate negative transfer effects. The CDNet [31] presents a novel approach for recognizing complex facial expressions across databases

with only a limited number of camera angles. By employing a cascaded decomposition network, the CDNet effectively handles various scenarios of facial expressions and achieves more accurate recognition performance in cross-domain learning tasks. On the other hand, Zhu et al. [18] introduced a method called multi-representation adaptation that focuses on aligning the distributions of multiple representations. On the basis of this approach, the Multi-Representation Adaptation Network (MRAN) can capture different aspects of image information to successfully accomplish cross-database facial expression recognition.

3. Multiple-Representation Joint Dynamic Domain Adaptive Network

3.1. Network Overview

The multi-representation joint dynamic domain adaptive network (MJDDAN) introduces global adaptive maximum mean difference (MMD) [24,32,33] and local subdomain adaptive LMMD [28,34,35] into the multi-representation feature extraction module, and it carries out joint dynamic domain adaptive in each substructure, respectively, to obtain fine-grained information about different feature spaces. The MJDDAN model framework is shown in Figure 1. The whole network structure can be divided into three parts. The first part is the ResNet network [36] model, represented by G, which is used to extract low-dimensional features of facial expression images. The second part is a multi-feature extraction module embedded with Joint Dynamic Maximum Mean Difference (JD-MMD), represented by H, which incorporates diverse substructures for extracting feature information from multiple perspectives. Within each substructure, JD-MMD combines a globally domain adaptive MMD module [24,32,33] and locally subdomain adaptive LMMD module [28,34,35], facilitating domain adaptation. The third part is classifier S.

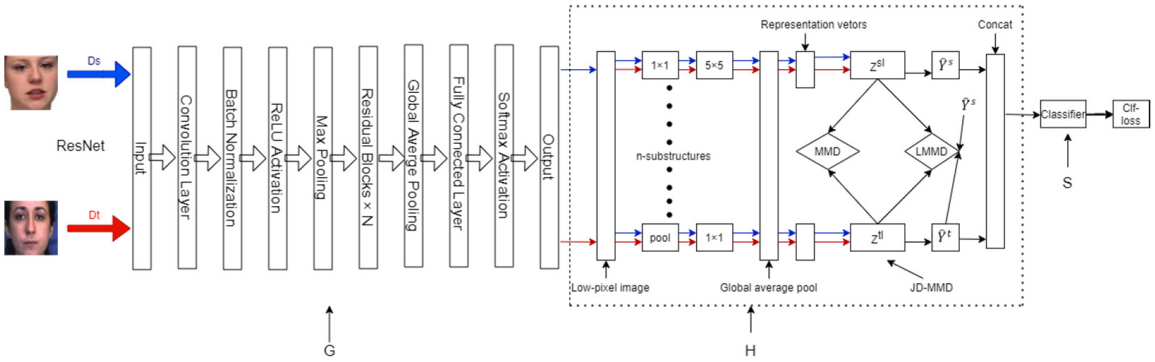


Figure 1. Structure diagram of MJDDAN. Low-pixel image represents low-dimensional feature extracted by ResNet [36] network, N-substructures represent multiple substructures, and Global average pool is global average pooling layer. Representation vectors are feature representations of different subspaces, Concat concatenates different feature representations, and JD-MMD is Joint Dynamic Maximum Mean Difference.

3.2. Multi-Representation Feature Extraction Module

In domain adaptive image classification, we usually use convolutional neural networks to extract image feature representations. However, due to the limitation of the network structure, the extracted features often contain only part of the information of the picture, which may lead to an unsatisfactory migration effect [18]. In Figure 2, (a) represents the original facial expression picture, while (b)–(d) represent the facial expression features after feature extraction. It can be observed that the facial expression features extracted only contain part of the picture's information [18]. These features may contain either useful information that is adaptive to subsequent domains or information that is not adaptive to the domain. In order to solve this problem, it is necessary to comprehensively consider the feature information of the original facial expression images from multiple perspectives when feature extraction is carried out. Through a comprehensive consideration of multiple perspectives, we can obtain a more comprehensive description of image features so as to

carry out a more comprehensive analysis and gain a better understanding of the original face expression image object. This holistic approach helps improve the effectiveness and accuracy of domain adaptation and reduces the possibility of extracting irrelevant or redundant information. To avoid the time consumption of training multiple convolutional neural networks, one option is to extract different features by introducing different components into the neural network structure [18].

Figure 2. Feature extraction from different angles.

The whole MJDDAN model can be divided into three modules: $g(\cdot)$ is the ResNet50 [36] network module used to extract features, $h(\cdot)$ is used to replace the average pooling layer with a multi-feature extraction module, and JD-MMD, $s(\cdot)$ is the classification module. The final complete network model is [16]

$$y = f(x) = (s \cdot h \cdot g)(x)(h \cdot g)(x) = h(g(x)) \tag{1}$$

The multi-representation feature extraction module has n distinct substructures, $h_1(\cdot) \cdots h_n(\cdot)$, each of which consists of two convolution operations or pooling operations. The objective optimization function of the multi-representation feature extraction module is as follows:

$$min \sum_{i}^{n} \hat{d}((h_i \cdot g)(X_s), (h_i \cdot g)(X_t)) \tag{2}$$

X_s indicates the source domain database, X_t indicates the target domain database, and $\hat{d}(\cdot, \cdot)$ indicates the distance between the source domain and the target domain [18].

After domain adaptive is completed in multiple subspaces, it is necessary to carry out the whole classification task. Multiple sub-feature extraction structures are connected through a full-join layer to obtain a join vector $[(h_1 \cdot g)(X), \cdots (h_n \cdot g)(X)]$, which is then input into the classifier $s(\cdot)$. Finally, the entire network model $y = f(x)$ embedded with the multi-representation feature extraction module can be expressed in the following form [18]

$$y = f(x) = s([(h_1 \cdot g)(X), \cdots (h_n \cdot g)(X)]). \tag{3}$$

3.3. Joint Dynamic Maximum Mean Difference

When unsupervised domain adaptation is carried out, the labeled samples are stored in the source domain as $D_S = \{x_i^s, y_i^s\}_{i=0}^{n_s}$, and the unlabeled samples are stored as $D_T = \{x_i^t\}_{i=0}^{n_t}$. n_s and n_t represent the number of samples in the source domain and the target domain, respectively, and y_i is used to indicate the label information of the I−th sample in the source domain. D_S and D_T are sampled from different data distributions, p and q, where p $\neq$ q. The purpose of transfer learning is to reduce the distribution difference between the source domain and the target domain by designing a network $f(x)$ [27,37–40].

In past attempts of unsupervised domain adaptation, the alignment between the source domain and target domain was in the global domain, but the alignment between the two domains was ignored, which not only means that the fine-grained information of the class was ignored, but also the discriminant structure became chaotic. Therefore, this paper proposes the Joint Dynamic Maximum Mean Difference (JD-MMD) to deal with

local subdomain adaptive as well as global adaptive. While minimizing the MMD, the LMMD is simultaneously used to introduce the target domain sample labels predicted by the network model into the feature alignment process so as to realize a joint dynamic domain adaptive method that focuses more on fine-grained information. The global MMD expression is [24,28,32,33]

$$MMD_H(P,Q) \triangleq || E_p[\varphi(x^s)] - E_q[\varphi(x^t)] ||_H^2 \tag{4}$$

The globally adaptive loss function L_M is denoted by $MMD_H(P,Q)$.
The local LMMD expression is [28,34,35]

$$LMMD_H(P,Q) \triangleq E_c \left|\left| E_{p^{(c)}}[\varphi(x^s)] - E_{q^{(c)}}[\varphi(x^t)] \right|\right|_H^2 \tag{5}$$

By minimizing the above formula, we can narrow the distance between subfields of the same class. In order to carry out domain adaptation more finely, an unbiased estimator is made according to the weight w^c of each class, which can evaluate the distance of the degree quantum field more accurately.

$$L\hat{M}D_H(P,Q) = \frac{1}{C}\sum_{c=1}^{C} || \sum_{x_i^s \in D_s} w_i^{sc}\varphi(x_i^s) - \sum_{x_j^t \in D_t} w_j^{tc}\varphi\left(x_j^t\right) ||_H^2 \tag{6}$$

In the given equation, the weights corresponding to x_i^s and x_j^t in category C are denoted as w_i^{sc} and w_j^{tc}, respectively. $\sum_{x_i^s \in D_s} w_i^{sc} = \sum_{x_j^t \in D_t} w_j^{tc} = 1$. The weighted sum of category C is represented as $\sum_{x_i \in D} w_i^c \varphi(x_i)$. The calculation formula for w_i^c is as follows [28]:

$$w_i^c = \frac{y_{ic}}{\sum_{(x_i,y_j)\in D} y_{jc}} \tag{7}$$

For the source domain, we can easily obtain the label y_{ic} but during unsupervised domain adaptation, it is difficult to obtain the label information in the target domain. In order to deal with the lack of labels in the target domain, the probability distribution output by the deep neural network can be used as the soft label of the target domain sample, and these soft labels can be used to calculate the difference between the distribution of related subdomains to achieve domain adaptation. So, the formula is [28]

$$\begin{aligned} \mathrm{L\hat{M}MD}_L(P,Q) = \; & \tfrac{1}{C}\textstyle\sum_{c=1}^{C}\Bigg[\sum_{i=1}^{n_s}\sum_{j=1}^{n_t} w_i^{sc}w_j^{sc}k\left(z_i^{sl},z_j^{sl}\right) \\ & + \sum_{i=1}^{n_s}\sum_{j=1}^{n_t} w_i^{tc}w_j^{tc}k\left(z_i^{tl},z_j^{tl}\right) - 2\sum_{i=1}^{n_s}\sum_{j=1}^{n_t} w_i^{sc}w_j^{tc}k\left(z_i^{sl},z_j^{tl}\right)\Bigg] \end{aligned} \tag{8}$$

where z^l is the L-layer activation $l \in \{1,2,\ldots,L\}$.

The local subdomain adaptation loss function L_L is denoted by $\mathrm{L\hat{M}MD}_L(P,Q)$, so joint dynamic domain adaptive can be expressed as

$$L_J = \lambda_M \times L_M + \lambda_L \times L_L \tag{9}$$

λ_M and λ_L are the weight parameters of the two domain adaptive methods, which are used to balance the importance of the two domain adaptive loss terms. L_J represents the loss function of joint dynamic domain adaptation, which realizes domain adaptation by dynamically adjusting the MMD and LMMD.

In order to prevent the network from falling into the optimal solution during training and to balance the optimization progress of two loss functions, L_M and L_L, we use a set of dynamic weighting factors for training. These dynamic weighting factors can be adjusted

according to the performance of the model during training to ensure a balance between the training speed of different loss functions. The dynamic weighting factor is expressed as follows:

$$\lambda_i = \frac{\alpha_i \times L_i}{\sum_{i \in \{M,L\}} L_i} \tag{10}$$

When training the network model, we use a fixed initial hyperparameter α_i to enhance the contribution of different loss functions. In order to achieve the global optimal solution, we introduce the dynamic weight factor α_i, which adjusts as the model iteratively updates the loss function value L_i. This dynamic adjustment can balance the contribution of different loss functions in the model training and ensure that the model does not fall into the local optimal solution to achieve the best migration effect.

$$\min_f \frac{1}{n_s} \sum_{i=1}^{n_s} J(f(x_i^s), y_i^s) + \lambda \sum_i^n \hat{d}_J((h_1 \cdot g)(X), \cdots (h_n \cdot g)(X)) \tag{11}$$

After feature alignment, the data from the source domain D_S and target domain D_T enter the classifier module S, but only the data from the source domain D_S undergoes classification loss training because the classification performance of the network model needs to be controlled by the labeled source domain D_S data, and then the parameters of the entire network model are adjusted during training. Therefore, the objective function of MJDDAN mainly includes cross entropy loss function and joint dynamic domain adaptive loss function. In the loss formula of MJDDAN, $J(\cdot, \cdot)$ is the cross-entropy loss function, that is, the classification loss; $\hat{d}_J(\cdot, \cdot)$ is the joint dynamic domain adaptive loss; L_J in the substructure is calculated using Equation (11); and $\lambda > 0$ is the tradeoff parameter.

4. Experimental Results and Analysis

The experimental environment and database employed in this study are detailed in this section. Within this experimental framework, we conduct a comparative analysis as well as an ablation experiment of the MJDDAN model for cross-database facial expression recognition.

4.1. Experimental Environment

The experimental setup of this study involves two distinct operating system platforms. On the personal host, we employ the Windows operating system for video processing in the database, to extract required images, and to conduct comparative experiments on traditional migration methods. On the server side, we conduct deep domain adaptive experiments using the Ubuntu operating system (graphics card: NVIDIA GTX1080Ti).

4.2. Facial Expression Database

The experiments in this paper encompass three distinct databases, providing a comprehensive validation of the proposed methodology. These databases include the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [19], the Bimodal Face and Body Gesture Database (FABO) [20], and the eNTERFACE '05 Audio-Visual Emotion Database (eNTERFACE) [21].

(1) The RAVDESS database is a multi-modal emotion database that includes two parts (voice emotions and song emotions). The performances of 12 male and 12 female professional actors were recorded, with each actor performing in a neutral American accent and ensuring correct content matching with different vocabulary. In the song mode, there are five types of emotions (tranquility, joy, grief, fear, and anger), while in the voice mode, there are expressions of serenity, joy, grief, anger, fear, surprise, and disgust. Each emotion has a neutral expression as well as two expressions of varying intensity [19,41–44]. Figure 3 is a sample of partial expressions in the RAVDESS database.

Figure 3. Sample of partial expressions in RAVDESS database.

(2) The FABO database was created by Hatice Gunes and others at the University of Technology Sydney in 2005, and it is the first publicly available bimodal emotion database that contains data on facial expressions and body language. Out of the 23 participants in the database, 11 are men, and 12 are women. During the process of collecting data, each subject was presented with different illustrations and situations, and their generated facial expressions and body language were recorded. Each video contains 2–4 fully displayed emotional expressions of the same type. The database recorded nine emotional expressions (surprise, anger, disgust, fear, boredom, anxiety, and uncertainty) [20,45–48]. Figure 4 is a sample of partial expressions in the FABO database.

Figure 4. Sample of partial expressions in FABO database.

(3) Martin and his team at KU Leuven developed the eNTERFACE database, which is a two-modal emotion database containing speech emotion and video emotion. The database was created with the selection of 46 subjects, 81% of whom were male, with the remaining being female, from 14 different countries. The participants expressed genuine feelings of happiness, fear, anger, surprise, disgust, and sadness in six different situations. Finally, the database was reviewed by a human expert who deleted the video data with unclear emotional expressions. In the end, data from 42 participants were retained [21,49–52]. Figure 5 is a sample of partial expressions in the eNTERFACE database.

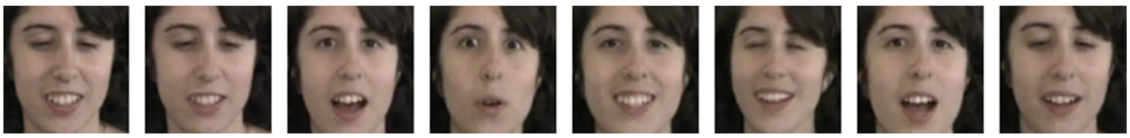


Figure 5. Sample of expressions in eNTERFACE database.

4.3. Experimental Details

As the three databases all store video information, it is necessary to extract image data containing emotional information from these videos. Initially, the OpenCV library is employed to process the videos into sequential frames, and subsequently, eight frames are selected at regular intervals for each video. Furthermore, considering that each database is collected in distinct environments, it becomes imperative to customize them by eliminating irrelevant data while preserving segments that have potential for generating emotional information. The objective of this customization process is to optimize the dataset which focuses on emotionally relevant content. Specifically, the facial region of interest spans from the left ear to the right ear and includes areas above (such as the forehead) and below it (including the chin), as depicted in Figure 6. This cropping technique aids in retaining crucial facial features for enhanced sentiment analysis and recognition.

Figure 6. Face clipping diagram.

To ensure the consistency of label information between the source and target domains, we selected anger, disgust, fear, happiness, sadness, and surprise as common emotion classification labels based on the criteria of the eNTERFACE database, which offers a minimal number of emotion classifications. By refining these three databases and standardizing their label information to encompass these six emotions, we can guarantee coherence and comparability in tasks related to emotion recognition.

Finally, in the RAVDESS database, we selected 192 video samples for each of the six emotions, resulting in a total of 1152 samples. Similarly, in the eNTERFACE database, we chose 215 video samples for each emotion, resulting in a total of 1290 video samples. In terms of the FABO dataset, varying quantities were selected for each of the six emotions (anger had 230 samples, disgust had 90 samples, fear had 86 samples, happiness had 118 samples, sadness had 68 samples, and surprise had 64 samples), resulting in a total of 656 samples. By employing such diverse sample selections across multiple databases, we can acquire an ample number of examples to effectively conduct experiments and evaluations on facial expression classification algorithms.

4.4. Cross-Database Facial Expression Recognition Experiment and Result Analysis

The experiment conducts nine transfer learning tasks on three databases, including $F \rightarrow R$, $e \rightarrow R$, $F\&e \rightarrow R$, $R \rightarrow F$, $e \rightarrow F$, $R\&e \rightarrow F$, $R \rightarrow e$, $F \rightarrow e$, and $R\&F \rightarrow e$. By comparing experiments and conducting ablation tests, the effectiveness of MJDDAN as well as the impact of each module on the overall network performance are demonstrated.

The existing transfer learning methods for cross-database facial expression recognition primarily encompass TCA [24], GFK [25], SA [26], DANN [27], DSAN [28], CDNet [31], and MRAN [18]. In this paper, we propose the MJDDAN model, which is compared with the above methods under identical experimental conditions. The evaluation criteria include the F1-Score, confusion matrix, and accuracy rate. The accuracy rates of the experimental results are presented in Table 1, while the F1-Scores are displayed in Table 2.

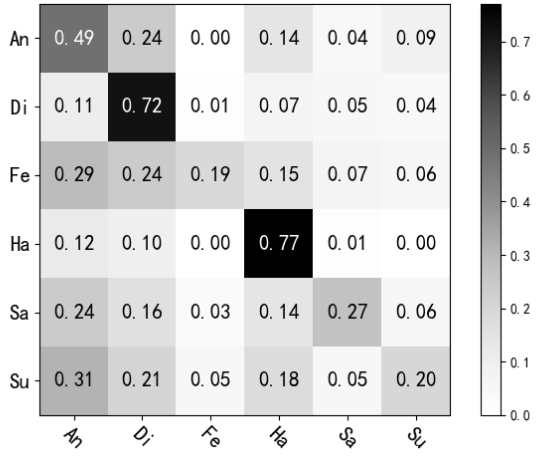
Table 1. The experimental accuracy of the proposed MJDDAN and other compared methods.

	TCA	GFK	SA	DANN	DSAN	CDNet	MRAN	MJDDAN
$F \rightarrow R$	31.47%	34.83%	35.65%	38.94%	40.75%	41.06%	43.27%	43.92%
$e \rightarrow R$	39.28%	42.39%	42.76%	44.27%	46.89%	48.66%	50.52%	52.53%
$F\&e \rightarrow R$	39.75%	43.27%	43.94%	46.63%	48.37%	51.85%	50.95%	53.64%
$R \rightarrow F$	28.35%	31.62%	31.82%	34.28%	35.89%	38.63%	38.10%	39.84%
$e \rightarrow F$	33.91%	37.83%	37.42%	39.16%	40.23%	42.72%	40.79%	42.54%
$R\&e \rightarrow F$	33.33%	38.42%	37.33%	40.49%	40.98%	42.81%	41.36%	43.66%
$R \rightarrow e$	27.83%	29.73%	30.03%	29.06%	31.67%	31.56%	30.92%	32.29%
$F \rightarrow e$	27.79%	30.92%	31.84%	32.23%	32.78%	33.37%	32.68%	33.96%
$R\&F \rightarrow e$	28.67%	31.85%	32.29%	32.03%	34.34%	34.85%	33.77%	35.87%

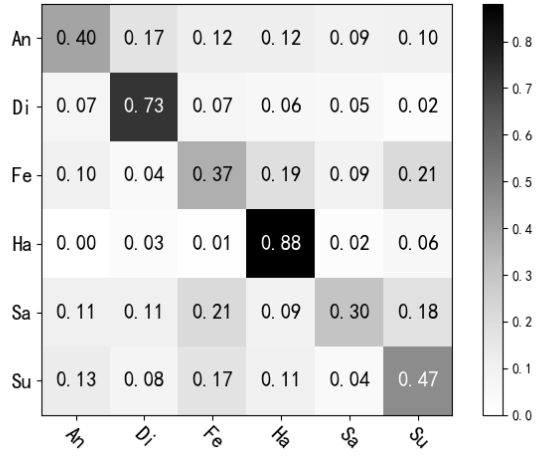
Table 2. The experimental F1-Scores of the proposed MJDDAN and other compared methods.

	TCA	GFK	SA	DANN	DSAN	CDNet	MRAN	MJDDAN
$F \rightarrow R$	0.3168	0.3462	0.3527	0.3940	0.4088	0.4127	0.4478	0.4672
$e \rightarrow R$	0.3953	0.4271	0.4288	0.4436	0.4657	0.4835	0.5004	0.5191
$F\&e \rightarrow R$	0.3992	0.4357	0.4425	0.4683	0.4851	0.5127	0.5063	0.5295
$R \rightarrow F$	0.2965	0.3133	0.3171	0.3491	0.3588	0.3759	0.3905	0.4028
$e \rightarrow F$	0.3477	0.3791	0.3769	0.3966	0.4016	0.4236	0.4153	0.4314
$R\&e \rightarrow F$	0.3448	0.3869	0.3745	0.4027	0.4077	0.4257	0.4285	0.4652
$R \rightarrow e$	0.2864	0.3016	0.3020	0.2973	0.3181	0.3282	0.3127	0.3250
$F \rightarrow e$	0.2839	0.3077	0.3154	0.3244	0.3292	0.3384	0.3239	0.3436
$R\&F \rightarrow e$	0.2894	0.3150	0.3251	0.3252	0.3450	0.3404	0.3356	0.3612

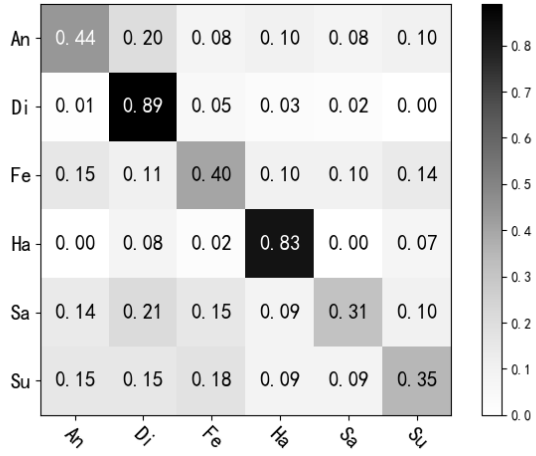
According to the experimental data presented in Table 1, the following conclusions can be drawn: The proposed MJDDAN outperforms traditional transfer learning methods (TCA, GFK, and SA) and deep learning methods (DANN, DSAN, CDNet, and MADTN) in cross-database facial expression recognition tasks. In Table 1, the bold text highlights the maximum value for each row's result. Generally speaking, deep learning methods exhibit superior performance compared to traditional approaches in this task. Traditional methods, such as TCA and SA, rely on linear transformations that are inadequate for nonlinear data mapping. Despite GFK being a nonlinear method that utilizes the locally linear embedding technique, it still fails to achieve satisfactory results. Therefore, traditional methods are not ideal for cross-database facial expression recognition tasks. Among the mentioned deep learning methods, DANN and DSAN only consider global or local domain adaptation, respectively, and both employ a single network structure for the domain adaptive deep learning approach. The proposed MJDDAN performs domain adaptation by simultaneously using four different substructures across four feature subspaces. This significantly enhances the accuracy of cross-database expression recognition. Compared with the latest CDNet method, only the experimental accuracy of the $e \rightarrow F$ cross-database expression recognition task is slightly lower, and the accuracy of the other recognition tasks is higher than that of the current excellent methods. Compared with the MRAN method, the proposed MJDDAN applies the local subdomain adaptive method in multiple subspaces to make use of more fine-grained information in different expression categories. According to the experimental results, the proposed MJDDAN in this paper achieves higher accuracy than the MRAN. For cross-database facial expression recognition tasks, the accuracy of the proposed MJDDAN is improved by up to 2.23% for the RAVDESS target domain. For the FABO target domain, the accuracy is improved by up to 1.54%. For the eNTERFACE target domain, the accuracy is improved by up to 1.76%. Among the nine groups of cross-database facial expression recognition experiments, the confusion matrix of the proposed MJDDAN is shown in Figure 7. An, Di, Fe, Ha, Sa, and Su represent anger, disgust, fear, happiness, sadness, and surprise, respectively.



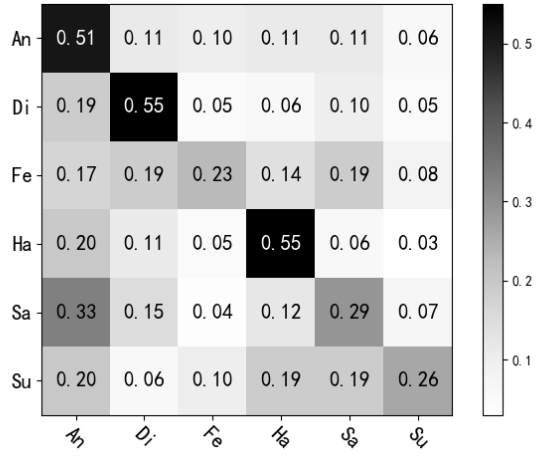
(**a**)



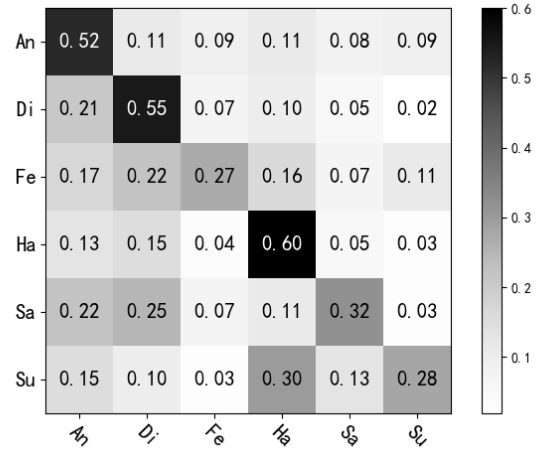
(**b**)



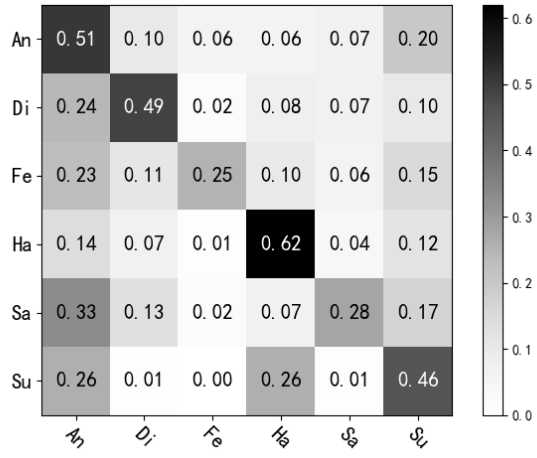
(**c**)



(**d**)



(**e**)



(**f**)

Figure 7. *Cont*.

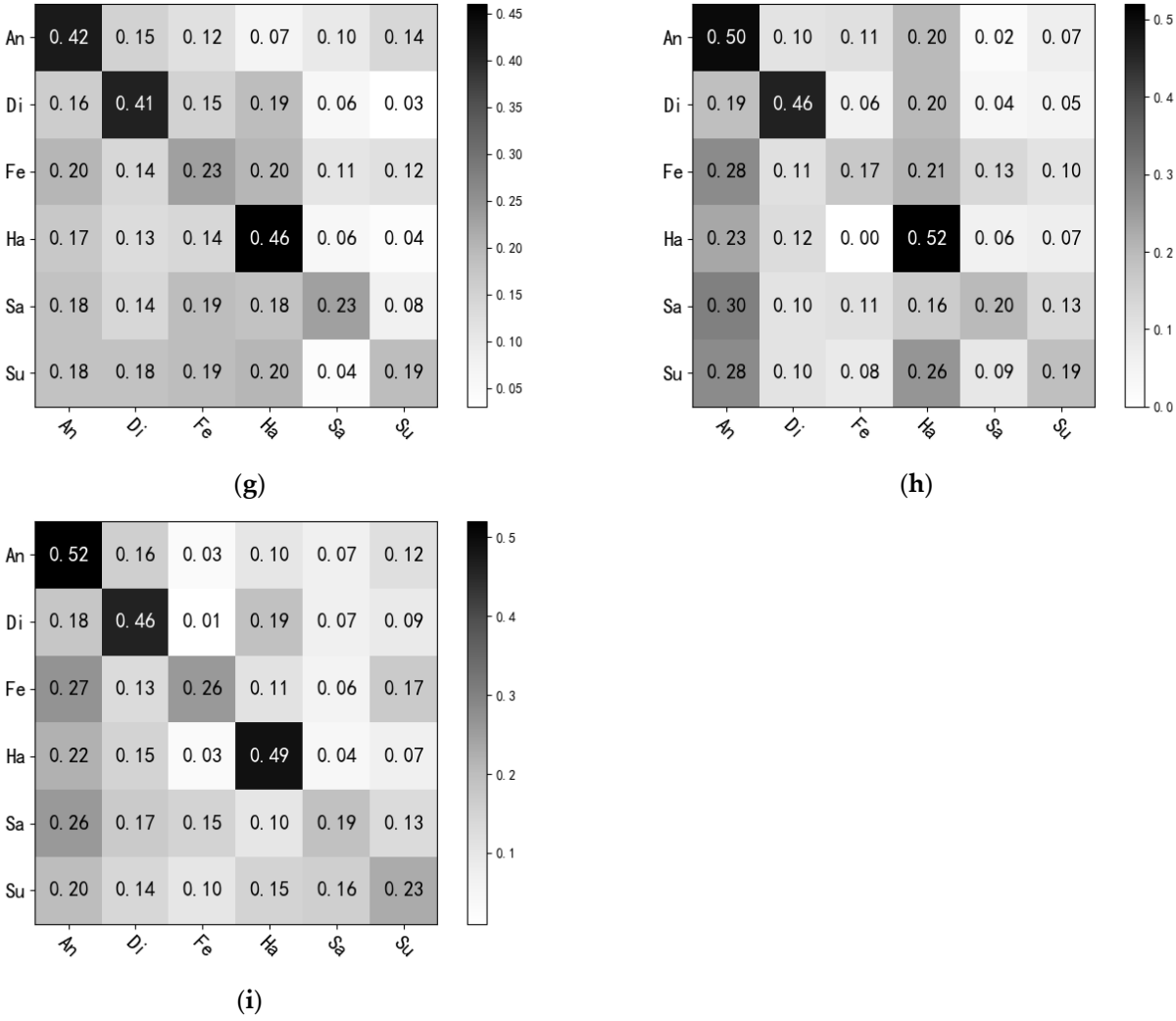


Figure 7. The experimental confusion matrix of the MJDDAN for nine groups of cross-database facial expression recognition experiments. (**a**) $F \rightarrow R$: The cross-database facial expression recognition task. (**b**) $e \rightarrow R$: The cross-database facial expression recognition task. (**c**) $F\&e \rightarrow R$: The cross-database facial expression recognition task. (**d**) $R \rightarrow F$: The cross-database facial expression recognition task. (**e**) $e \rightarrow F$: The cross-database facial expression recognition task. (**f**) $R\&e \rightarrow F$: The cross-database facial expression recognition task. (**g**) $R \rightarrow e$: The cross-database facial expression recognition task. (**h**) $F \rightarrow e$: The cross-database facial expression recognition task. (**i**) $R\&F \rightarrow e$: The cross-database facial expression recognition task.

4.5. Ablation Experiment

Ablation experiments are conducted in this section to evaluate the contribution of the multi-representation feature extraction module and the JD-MMD module to the entire MJDDAN model. To ensure the effectiveness and fairness of the experiment, we use the same experimental environment and strategy as those used in the comparative experiment in the previous section. The RAVDESS database is used as the target domain data, while the FABO database and eNTERFACE data are used as the source domain data. We designed several variants of the network model to prove the effectiveness of each module. The first variant only adds the multi-representation feature extraction module to the ResNet network (MFER), without embedding the JD-MMD, in order to verify its contribution. The second method embeds the JD-MMD into the ResNet network (JDDAN) and performs joint dynamic domain adaptation for verification purposes. The third variant simultaneously adds both modules mentioned above to the ResNet network, creating the MJDDAN model. Moreover, a global adaptive network based on multi-representation feature extraction

(MGDAN) and a local subdomain adaptive network based on multi-representation feature extraction (MLDAN) are constructed to verify their effectiveness, respectively. Table 3 shows the ablation experiment accuracy rates and F1-Scores of the cross-database facial expression recognition task with the RAVDESS database as the target domain.

Table 3. Ablation experiment accuracy rates and F1-Scores of cross-database facial expression recognition task with RAVDESS database as target domain.

	F→*R*		*e*→*R*		*F*&*e*→*R*	
	ACC	**F1-Score**	**ACC**	**F1-Score**	**ACC**	**F1-Score**
ResNet	26.74%	0.2693	34.66%	0.3417	35.62%	0.3451
MFER	28.59%	0.2945	36.43%	0.3629	37.92%	0.3751
JDDAN	40.75%	0.4088	46.89%	0.4657	48.37%	0.4851
MGDAN	41.66%	0.4281	50.47%	0.5017	51.03%	0.5152
MLDAN	43.74%	0.4625	52.25%	0.5173	53.27%	0.5228
MJDDAN	43.92%	0.4672	52.53%	0.5191	53.64%	0.5295

The results of the ablation experiment presented in Table 3 demonstrate that the two modules contribute differently to cross-database facial expression recognition (the bold text indicates the maximum value for each line). Compared to the ResNet network based on multi-representation feature extraction, a multi-representation feature extraction module replaces the last fully connected layer. The experimental results show that all three cross-database facial expression recognition tasks achieve higher accuracy than the ResNet network, with improvements of 1.85%, 1.77%, and 2.3%, respectively, proving that the multi-representation feature extraction module can enhance the experimental accuracy without requiring domain adaptation between the source and target domains for cross-database facial expression recognition purposes. Furthermore, incorporating the JD-MMD into the last fully connected layer significantly reduces facial expression feature distribution across different databases compared to the ResNet network based on joint dynamic domain adaptation. As observed in the experimental results, all three cross-database facial expression recognition tasks yield substantially higher accuracy rates compared to both the ResNet network and multi-representation feature extraction network model, achieving increases of 14.01%, 12.23%, and 17.65%, respectively, compared to the ResNet network. In comparison to the ResNet network that relies on multi-representation feature extraction, our proposed MJDDAN demonstrates significant improvements of 12.16%, 10.46%, and 10.45%, respectively, in the cross-database facial expression recognition experiment, highlighting the importance of domain adaptation between different databases. By incorporating the MJDDAN into various subspaces of the ResNet network for multi-representation feature extraction, we effectively align feature distributions from diverse facial expression spaces to capture more detailed information. The experimental accuracy of all three cross-database facial expression recognition tasks increased by 3.17%, 5.64%, and 5.27%, respectively, when compared to the accuracy obtained when using a single network structure like ResNet with a joint dynamic domain adaptive network alone. These results indicate that the migration effect achieved by our multi-subspace joint dynamic domain adaptive method is significantly superior to that obtained with a single network structure. The proposed MJDDAN in this paper can enhance accuracy compared to both the global domain adaptive network and the local subdomain adaptive network based on multi-representation feature extraction. This finding proves that the proposed MJDDAN exhibits superior migration performance over both the global domain adaptive and local subdomain adaptive approaches.

The primary objective of this study is to address the issue of distribution discrepancy in cross-database facial expression recognition. The proposed approach in this paper involves extracting features from images through multiple sub-feature spaces at different angles and aligning the source and target domain features through the Joint Dynamic Maximum Mean Difference (JD-MMD) within the sub-feature space. The comparative

experiments revealed that, although the accuracy for the $e \rightarrow F$ experiment and the F1-Score for the $R \rightarrow e$ experiment are slightly lower than those of the CDNet method, our method outperforms other approaches in other respective experiments; therefore, the effectiveness of our proposed approach is partially validated.

5. Conclusions

This paper proposes a novel multi-representation joint dynamic domain adaptation network (MJDDAN) model for cross-database facial expression recognition. The model proposes a Joint Dynamic Maximum Mean Difference (JD-MMD) module and integrates it into each sub-feature space to effectively reduce the feature distribution differences between different domains. The computational cost of the JD-MMD module is slightly higher compared to that of the LMMD and MMD. Finally, extensive experiments are conducted using the MJDDAN model for cross-database facial expression recognition on the RAVDESS database, FABO database, and eNTERFACE database. Compared to the best comparison method chosen in this article, the accuracy of the MJDDAN model was improved by 1.79%, 0.85%, and 1.02%, respectively. In the future, we will apply the MJDDAN model to achieve large-scale cross-database facial expression recognition. Although the JD-MMD is an effective domain adaptation method, it may not necessarily be the best choice. In future research, we can design a network structure that can autonomously select domain adaptation methods that are suitable for each subspace and select the best measurement method through network training.

Author Contributions: Conceptualization, J.Y.; methodology, J.Y.; software, Y.Y. (Yuebo Yue); validation, J.Y., Y.Y. (Yuebo Yue) and K.Y.; formal analysis, X.Z.; investigation, Y.Y. (Yuebo Yue); data curation, K.Y.; writing—original draft preparation, Y.Y. (Yuebo Yue), J.W. and K.Y.; writing—review and editing X.Z. and Y.L.; supervision, Y.Y. (Yuebo Yue) and Y.Y. (Yuan Yang); project administration, J.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work is partly supported by the National Natural Science Foundation of China (NSFC) under Grants 61971236, is partly supported by Open Project of Blockchain Technology and Data Security Key Laboratory Ministry of Industry and Information Technology under Grants 20242218, Is partly supported by Natural Science Research Start up Foundation of Recruiting Talents of Nan[1]jing University of Posts and Telecommunications (Grant No.NY223030), is partly by Nanjing Science and Technology Innovation Foundation for Overseas Students under Grants NJKCZYZZ2023-04.

Institutional Review Board Statement: The study did not require ethical approval.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Experiments used publicly available datasets. The link for the FABO dataset is https://www.cl.cam.ac.uk/~hg410/fabo.html/, the link for the eNTERFACE dataset is https://www.enterface.net/enterface05/, the link for the RAVDESS dataset is https://zenodo.org/records/1188976/.

Conflicts of Interest: Author Xiaoyang Zhou was employed by the company China Mobile Zijin (Jiangsu) Innovation Research Institute Co., Ltd. Author Ying Liu was employed by the company China Mobile Communications Group Jiangsu Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Chen, W.; Wang, A. Enhanced Facial Expression Recognition Based on Facial Action Unit Intensity and Region. In Proceedings of the 2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Honolulu, HI, USA, 1–4 October 2023; pp. 1939–1944.
2. De, A.; Saha, A. A Comparative Study on Different Approaches of Real Time Human Emotion Recognition Based on Facial Expression Detection. In Proceedings of the 2015 International Conference on Advances in Computer Engineering and Applications, Ghaziabad, India, 19–20 March 2015; pp. 483–487.

3. Lin, S.-Y.; Tseng, Y.-W.; Wu, C.-R.; Kung, Y.-C.; Chen, Y.-Z.; Wu, C.-M. A Continuous Facial Expression Recognition Model Based on Deep Learning Method. In Proceedings of the 2019 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), Taipei, Taiwan, 3–6 December 2019; pp. 1–2.
4. Verma, K.; Khunteta, A. Facial Expression Recognition Using Gabor Filter and Multi-Layer Artificial Neural Network. In Proceedings of the 2017 International Conference on Information, Communication, Instrumentation and Control (ICICIC), Indore, India, 17–19 August 2017; pp. 1–5.
5. Wei, H.; Zhang, Z. A Survey of Facial Expression Recognition Based on Deep Learning. In Proceedings of the 2020 15th IEEE Conference on Industrial Electronics and Applications (ICIEA), Kristiansand, Norway, 9–13 November 2020; pp. 90–94.
6. Grover, R.; Bansal, S. Facial Expression Recognition: Deep Survey, Progression and Future Perspective. In Proceedings of the 2023 International Conference on Advancement in Computation & Computer Technologies (InCACCT), Gharuan, India, 5–6 May 2023; pp. 111–117.
7. Singh, Y.B.; Goel, S. Survey on Human Emotion Recognition: Speech Database, Features and Classification. In Proceedings of the 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), Greater Noida, India, 12–13 October 2018; pp. 298–301.
8. Li, Y.; Chao, L.; Liu, Y.; Bao, W.; Tao, J. From Simulated Speech to Natural Speech, What Are the Robust Features for Emotion Recognition? In Proceedings of the 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), Xi'an, China, 21–24 September 2015; pp. 368–373.
9. Luqin, S. A Survey of Facial Expression Recognition Based on Convolutional Neural Network. In Proceedings of the 2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS), Beijing, China, 17–19 June 2019; pp. 1–6.
10. Ming, F.J.; Shabana Anhum, S.; Islam, S.; Keoy, K.H. Facial Emotion Recognition System for Mental Stress Detection among University Students. In Proceedings of the 2023 3rd International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME), Tenerife, Spain, 19–21 July 2023; pp. 1–6.
11. Dixit, A.N.; Kasbe, T. A Survey on Facial Expression Recognition Using Machine Learning Techniques. In Proceedings of the 2nd International Conference on Data, Engineering and Applications (IDEA), Bhopal, India, 28–29 February 2020; pp. 1–6.
12. Fatjriyati Anas, L.; Ramadijanti, N.; Basuki, A. Implementation of Facial Expression Recognition System for Selecting Fashion Item Based on Like and Dislike Expression. In Proceedings of the 2018 International Electronics Symposium on Knowledge Creation and Intelligent Computing (IES-KCIC), Bali, Indonesia, 29–30 October 2018; pp. 74–78.
13. Tang, H.; Cen, X. A Survey of Transfer Learning Applied in Medical Image Recognition. In Proceedings of the 2021 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA), Dalian, China, 27–28 August 2021; pp. 94–97.
14. Kusunose, T.; Kang, X.; Kiuchi, K.; Nishimura, R.; Sasayama, M.; Matsumoto, K. Facial Expression Emotion Recognition Based on Transfer Learning and Generative Model. In Proceedings of the 2022 8th International Conference on Systems and Informatics (ICSAI), Kunming, China, 10–12 December 2022; pp. 1–6.
15. Bousaid, R.; El Hajji, M.; Es-Saady, Y. Facial Emotions Recognition Using Vit and Transfer Learning. In Proceedings of the 2022 5th International Conference on Advanced Communication Technologies and Networking (CommNet), Marrakech, Morocco, 12–14 December 2022; pp. 1–6.
16. Zhou, J.; Zhuang, J.; Li, B.; Zhou, L. Research on Underwater Image Recognition Based on Transfer Learning. In Proceedings of the OCEANS 2022, Hampton Roads, Virginia Beach, VA, USA, 17–20 October 2022; pp. 1–7.
17. Xia, K.; Gu, X.; Chen, B. Cross-Dataset Transfer Driver Expression Recognition via Global Discriminative and Local Structure Knowledge Exploitation in Shared Projection Subspace. *IEEE Trans. Intell. Transp. Syst.* **2021**, *22*, 1765–1776. [CrossRef]
18. Zhu, Y.; Zhuang, F.; Wang, J.; Chen, J.; Shi, Z.; Wu, W.; He, Q. Multi-Representation Adaptation Network for Cross-Domain Image Classification. *Neural Netw.* **2019**, *119*, 214–221. [CrossRef] [PubMed]
19. Livingstone, S.R.; Russo, F.A. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A Dynamic, Multimodal Set of Facial and Vocal Expressions in North American English. *PLoS ONE* **2018**, *13*, e0196391. [CrossRef] [PubMed]
20. Gunes, H.; Piccardi, M. A Bimodal Face and Body Gesture Database for Automatic Analysis of Human Nonverbal Affective Behavior. In Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06), Hong Kong, China, 20–24 August 2006; pp. 1148–1153.
21. Martin, O.; Kotsia, I.; Macq, B.; Pitas, I. The ENTERFACE'05 Audio-Visual Emotion Database. In Proceedings of the 22nd International Conference on Data Engineering Workshops (ICDEW'06), Atlanta, GA, USA, 3–7 April 2006; p. 8.
22. Ni, T.; Zhang, C.; Gu, X. Transfer Model Collaborating Metric Learning and Dictionary Learning for Cross-Domain Facial Expression Recognition. *IEEE Trans. Comput. Soc. Syst.* **2021**, *8*, 1213–1222. [CrossRef]
23. Zhang, W.; Song, P.; Zheng, W. Joint Local-Global Discriminative Subspace Transfer Learning for Facial Expression Recognition. *IEEE Trans. Affect. Comput.* **2023**, *14*, 2484–2495. [CrossRef]
24. Pan, S.J.; Tsang, I.W.; Kwok, J.T.; Yang, Q. Domain Adaptation via Transfer Component Analysis. *IEEE Trans. Neural Netw.* **2011**, *22*, 199–210. [CrossRef] [PubMed]
25. Gong, B.; Shi, Y.; Sha, F.; Grauman, K. Geodesic Flow Kernel for Unsupervised Domain Adaptation. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 2066–2073.
26. Fernando, B.; Habrard, A.; Sebban, M.; Tuytelaars, T. Unsupervised Visual Domain Adaptation Using Subspace Alignment. In Proceedings of the 2013 IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 2960–2967.

27. Ganin, Y.; Lempitsky, V. Unsupervised Domain Adaptation by Backpropagation. In Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 7–9 July 2015; PMLR: Cambridge, MA, USA, 2015; Volume 37, pp. 1180–1189.
28. Zhu, Y.; Zhuang, F.; Wang, J.; Ke, G.; Chen, J.; Bian, J.; Xiong, H.; He, Q. Deep Subdomain Adaptation Network for Image Classification. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *32*, 1713–1722. [CrossRef] [PubMed]
29. Ren, C.-X.; Dai, D.-Q.; Huang, K.-K.; Lai, Z.-R. Transfer Learning of Structured Representation for Face Recognition. *IEEE Trans. Image Process.* **2014**, *23*, 5440–5454. [CrossRef] [PubMed]
30. Long, M.; Wang, J.; Ding, G.; Shen, D.; Yang, Q. Transfer Learning with Graph Co-Regularization. *IEEE Trans. Knowl. Data Eng.* **2014**, *26*, 1805–1818. [CrossRef]
31. Zou, X.; Yan, Y.; Xue, J.H.; Chen, S.; Wang, H. Learn-to-Decompose: Cascaded Decomposition Network for Cross-Domain Few-Shot Facial Expression Recognition. In Proceedings of the Computer Vision—ECCV 2022, Tel Aviv, Israel, 23–27 October 2022; Springer Nature: Cham, Switzerland, 2022; pp. 683–700.
32. Zhang, W.; Wu, D. Discriminative Joint Probability Maximum Mean Discrepancy (DJP-MMD) for Domain Adaptation. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020; pp. 1–8.
33. Lin, W.; Mak, M.-M.; Li, N.; Su, D.; Yu, D. Multi-Level Deep Neural Network Adaptation for Speaker Verification Using MMD and Consistency Regularization. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6839–6843.
34. Zhang, Z.; Wu, D.; Zhu, D.; Zhang, X. Ground Moving Target Detection for Multichannel SAR System Based on Subdomain Adaptation. In Proceedings of the 2023 IEEE 23rd International Conference on Communication Technology (ICCT), Wuxi, China, 20–22 October 2023; pp. 156–161.
35. Xiao, H.; Dong, L.; Wang, W.; Ogai, H. Distribution Sub-Domain Adaptation Deep Transfer Learning Method for Bridge Structure Damage Diagnosis Using Unlabeled Data. *IEEE Sens. J.* **2022**, *22*, 15258–15272. [CrossRef]
36. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
37. Zhang, W.; Wang, J.; Wang, Y.; Wang, F.-Y. ParaUDA: Invariant Feature Learning With Auxiliary Synthetic Samples for Unsupervised Domain Adaptation. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 20217–20229. [CrossRef]
38. Xie, J.; Zhou, Y.; Xu, X.; Wang, G.; Shen, F.; Yang, Y. Region-Aware Semantic Consistency for Unsupervised Domain-Adaptive Semantic Segmentation. In Proceedings of the 2023 IEEE International Conference on Multimedia and Expo (ICME), Brisbane, Australia, 10–14 July 2023; pp. 90–95.
39. Gan, B.; Dong, Q. Unsupervised Domain-Adaptive Image Classification Algorithm Incorporating Generative Adversarial Networks. In Proceedings of the 2021 3rd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), Taiyuan, China, 3–5 December 2021; pp. 305–310.
40. Chen, Y.; Yang, C.; Zhang, Y.; Li, Y. Conditional Adaptation Deep Networks for Unsupervised Cross Domain Image Classifcation. In Proceedings of the 2019 14th IEEE Conference on Industrial Electronics and Applications (ICIEA), Xi'an, China, 19–21 June 2019; pp. 517–521.
41. Rochlani, Y.R.; Raut, A.B. Machine Learning Approach for Detection of Speech Emotions for RAVDESS Audio Dataset. In Proceedings of the 2024 Fourth International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), Bhilai, India, 11–12 January 2024; pp. 1–7.
42. Agrima, A.; Barakat, A.; Mounir, I.; Farchi, A.; ElMazouzi, L.; Mounir, B. Speech Emotion Recognition Using Energies in Six Bands and Multilayer Perceptron on RAVDESS Dataset. In Proceedings of the 2022 5th International Conference on Advanced Communication Technologies and Networking (CommNet), Marrakech, Morocco, 12–14 December 2022; pp. 1–5.
43. Sowmya, G.; Naresh, K.; Sri, J.D.; Sai, K.P.; Indira, D.N.V.S.L.S. Speech2Emotion: Intensifying Emotion Detection Using MLP through RAVDESS Dataset. In Proceedings of the 2022 International Conference on Electronics and Renewable Systems (ICEARS), Tuticorin, India, 16–18 March 2022; pp. 1–3.
44. Singh, V.; Prasad, S. Speech Emotion Recognition Using Fully Convolutional Network and Augmented RAVDESS Dataset. In Proceedings of the 2023 International Conference on Advanced Computing Technologies and Applications (ICACTA), Bali, Indonesia, 5–6 October 2023; pp. 1–7.
45. Noroozi, F.; Corneanu, C.A.; Kaminska, D.; Sapinski, T.; Escalera, S.; Anbarjafari, G. Survey on Emotional Body Gesture Recognition. *IEEE Trans. Affect. Comput.* **2021**, *12*, 505–523. [CrossRef]
46. Gunes, H.; Piccardi, M. Creating and Annotating Affect Databases from Face and Body Display: A Contemporary Survey. In Proceedings of the 2006 IEEE International Conference on Systems, Man and Cybernetics, Taipei, Taiwan, 8–11 October 2006; pp. 2426–2433.
47. Alepis, E.; Stathopoulou, I.-O.; Virvou, M.; Tsihrintzis, G.A.; Kabassi, K. Audio-Lingual and Visual-Facial Emotion Recognition: Towards a Bi-Modal Interaction System. In Proceedings of the 2010 22nd IEEE International Conference on Tools with Artificial Intelligence, Arras, France, 27–29 October 2010; pp. 274–281.
48. Chen, S.; Tian, Y.; Liu, Q.; Metaxas, D.N. Segment and Recognize Expression Phase by Fusion of Motion Area and Neutral Divergence Features. In Proceedings of the Face and Gesture 2011, Santa Barbara, CA, USA, 21–23 March 2011; pp. 330–335.
49. Chen, L.; Wang, K.; Li, M.; Wu, M.; Pedrycz, W.; Hirota, K. *K*-Means Clustering-Based Kernel Canonical Correlation Analysis for Multimodal Emotion Recognition in Human–Robot Interaction. *IEEE Trans. Ind. Electron.* **2023**, *70*, 1016–1024. [CrossRef]

50. Li, L.; Zhao, Y.; Jiang, D.; Zhang, Y.; Wang, F.; Gonzalez, I.; Valentin, E.; Sahli, H. Hybrid Deep Neural Network--Hidden Markov Model (DNN-HMM) Based Speech Emotion Recognition. In Proceedings of the 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, Geneva, Switzerland, 2–5 September 2013; pp. 312–317.
51. Dhoot, A.; Hadj-Alouane, N.B.; Turki-Hadj Alouane, M. 2D CNN vs 3D CNN: An Empirical Study on Deep Learning-Based Facial Emotion Recognition. In Proceedings of the 2023 International Conference on Modeling, Simulation & Intelligent Computing (MoSICom), Dubai, UAE, 7–9 December 2023; pp. 138–143.
52. Toledo-Ronen, O.; Sorin, A. Voice-Based Sadness and Anger Recognition with Cross-Corpora Evaluation. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 7517–7521.