

Improved Cross-Dataset Facial Expression Recognition by Handling Data Imbalance and Feature Confusion

Manogna Sreenivas¹, Sawa Takamuku², Soma Biswas¹, Aditya Chepuri³
Balasubramanian Vengatesan³, and Naotake Natori²

¹ Indian Institute of Science, Bangalore, India

² Aisin Corporation, Japan

³ Aisin Automotive Haryana Pvt Ltd., Bangalore

Abstract. Facial Expression Recognition (FER) models trained on one dataset (source) usually do not perform well on a different dataset (target) due to the implicit domain shift between different datasets. In addition, FER data is naturally highly imbalanced, with a majority of the samples belonging to few expressions like neutral, happy and relatively fewer samples coming from expressions like disgust, fear, etc., which makes the FER task even more challenging. This class imbalance of the source and target data (which may be different), along with other factors like similarity of few expressions, etc., can result in unsatisfactory target classification performance due to confusion between the different classes. In this work, we propose an integrated module, termed **DIFC**, which can not only handle the source **D**ata **I**mbalance, but also the **F**eature **C**onfusion of the target data for improved classification of the target expressions. We integrate this DIFC module with an existing Unsupervised Domain Adaptation (UDA) approach to handle the domain shift and show that the proposed simple yet effective module can result in significant performance improvement on four benchmark datasets for Cross-Dataset FER (CD-FER) task. We also show that the proposed module works across different architectures and can be used with other UDA baselines to further boost their performance.

Keywords: Facial Expression Recognition, Unsupervised Domain Adaptation, Class Imbalance

1 Introduction

The need for accurate facial expression recognition models is ever increasing, considering its applications in driver assistance systems, understanding social behaviour in different environments [28, 8], Human Computer Interaction [10], security applications [1], etc. Though deep learning has proven to work remarkably well for generic object classification tasks, classifying human expressions still remains challenging due to the subjective nature of the task [16, 3]. Since

data annotation is itself an expensive task, it is important that FER models trained on one dataset can be adapted seamlessly to other datasets. The large domain shift between the real-world datasets [21, 25, 7, 13, 35] makes this problem extremely challenging. This problem is addressed in the UDA setting, where the objective is to transfer discriminative knowledge learnt from a labelled source domain to a target domain with different data distribution, from which only unlabelled samples are accessible. Most of the current UDA approaches [11, 31, 27] try to address this domain shift by aligning the source and target features.

In this work, we aim to address complementary challenges which can adversely affect the FER performance. One such challenge is the huge source data imbalance that is usually present in FER datasets since the number of labelled training examples for a few expressions like neutral and happy is usually significantly more compared to that of the less frequently occurring expressions like disgust, fear, etc. However, the final goal is to recognize the expressions in the target dataset, which may not have the same amount of imbalance as the source. In addition, there may be other factors, like the inherent similarity between two expressions etc. that may result in confusion between different target classes.

Here, we propose a simple, yet effective module termed as **DIFC** (**D**ata **I**mbalance and **F**eature **C**onfusion), which addresses both these challenges simultaneously. Specifically, we modify the state-of-the-art technique LDAM [4] which is designed to handle class imbalance for the supervised classification task, such that it can also handle target feature confusion in an unsupervised domain adaptation scenario. This module can be seamlessly integrated with several UDA methods [11, 27, 31]. In this work, we integrate it with an existing UDA technique, Maximum Classifier Discrepancy (MCD) [27] and evaluate its effectiveness for the CD-FER task on four benchmark datasets, namely JAFFE [25], SFEW2.0 [7], FER 2013 [13] and ExpW [35]. Thus, the contributions of this work can be summarized as:

1. We propose a simple yet effective DIFC module for the CD-FER task.
2. The module can not only handle source data imbalance, but also target feature confusion in an integrated manner.
3. The proposed module can be used with several existing UDA approaches for handling these challenges.
4. Experiments on four real-world benchmark datasets show the effectiveness of the proposed DIFC module.

2 Related Work

Unsupervised Domain Adaptation: The objective of UDA is to utilize labelled source domain samples along with unlabelled target domain samples to learn a classifier that can perform well on the target domain. A wide spectrum of UDA methods follow adversarial training mechanism inspired by Generative Adversarial Networks [12]. Domain Adversarial Neural Networks [11] uses a two-player game, where the feature extractor aims to align the source and target features, while the domain discriminator aims to distinguish between the domains.

This results in a domain invariant feature space and hence, translating the source classifier to the target. Conditional Adversarial Domain Adaptation [23] leverages the discriminative information provided by the classifier to aid domain alignment. Another line of works [15, 26] perform distribution alignment in the pixel space instead of feature space. Maximum Classifier Discrepancy (MCD) [27] uses two classifiers to identify misaligned target features and further aligns them in an adversarial fashion. In MDD [34], a new distribution divergence measure called Margin Disparity Discrepancy is proposed, which has a strong generalization bound that can ease minimax optimization. In SWD [17], the authors use a Wasserstein distance metric to measure the discrepancy between classifiers. SAFN [31] is a simple non-adversarial UDA method where the features are learnt progressively with large norms as they are better transferable to the target domain.

Cross-Dataset FER: Domain shifts in facial expression recognition can be attributed to several factors like pose, occlusion, race, lighting conditions etc. Also, inconsistencies in data due to noisy source labels, high class imbalance, etc., add to the challenges. We briefly discuss prior works [36, 32, 29, 20, 5] that address this problem. In [32], unsupervised domain adaptive dictionary learning is used to minimize the discrepancy between the source and target domains. A transductive transfer subspace learning method that combines the labelled source domain with an auxiliary set of unlabelled samples from the target domain to jointly learn a discriminative subspace was proposed in [36]. In [29], after pretraining the network using source samples, a Generative Adversarial Network is trained to generate image samples from target distribution, which is then used along with the source samples to finetune the network. Contrary to these works where only the marginal distribution between source and target are aligned, [20] proposes to minimize the discrepancy between the class conditional distributions across FER datasets as well. In [37], they propose to preserve class-level semantics while adversarially aligning source and target features, along with an auxiliary uncertainty regularization to mitigate the effect of uncertain images in CD-FER task. The recent state-of-the-art AGRA framework [30, 5] proposes to integrate graph representation adaptation with adversarial learning to adapt global and local features across domains effectively.

Class Imbalance Methods: The problem of class imbalance has been quite extensively studied in the supervised image classification setting. The traditional ways of mitigating imbalance are re-sampling and re-weighting. The majority classes are under-sampled while minority classes are over-sampled to balance the classes in a re-sampling strategy [2]. However, this duplicates the minority classes, leading to overfitting, while undersampling the majority classes could leave out important samples. On the other hand, re-weighting refers to adaptively weighing a sample’s loss based on its class. Weighting by inverse class frequency is a commonly used strategy. [6] propose to use a re-weighting scheme by estimating the effective number of samples per class. In [14], a hybrid loss

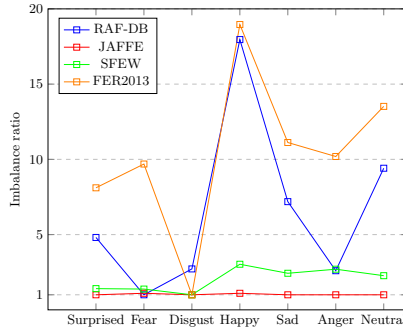


Fig. 1: Imbalance ratio for different classes vary across FER datasets.

is proposed that jointly performs classification and clustering. A max-margin constraint is used to achieve equispaced classification boundaries by simultaneously minimizing intra-class variance and maximizing inter-class variance. In LDAM [4], class imbalance is handled by enforcing a larger margin for the minority classes. They show that enforcing class-dependent margins along with a deferred re-weighting strategy effectively handles class imbalance in the supervised classification problem. We propose to modify this LDAM loss for the UDA setting to simultaneously handle source data imbalance and target confusion.

3 Problem Definition and Notations

We assume that we have access to a labelled source dataset $\mathcal{D}_s = \{(x_s^i, y_s^i)\}_{i=1}^{n_s}$, where $y_s^i \in \{1, \dots, K\}$ and K denotes the number of classes. The objective is to correctly classify the samples from the target dataset $\mathcal{D}_t = \{x_t^i\}_{i=1}^{n_t}$, which does not have the class labels, but shares the same label space as \mathcal{D}_s . Here, n_s and n_t denote the number of source and target samples respectively. Now, we discuss the proposed approach in detail.

4 Proposed Approach

Most of the existing UDA approaches [11, 27, 34, 23] aim to align the source and target features to mitigate the domain shift between them. In this work, we address two complementary challenges, namely source data imbalance and target feature confusion, which have been relatively less explored in the context of CD-FER task. These two challenges are handled in an integrated manner in the proposed DIFC module, which we describe next.

4.1 Handling Data Imbalance in DIFC

As already mentioned, FER datasets are naturally very imbalanced, with few classes like happy and neutral having a large number of samples, compared to

classes like fear, disgust, etc. This can be seen from Fig. 1, where we plot the imbalance ratio of all the classes for the four datasets used in this work.

For a given dataset, if we denote the number of samples available for class k by N_k , its imbalance ratio is calculated as

$$Imbalance_ratio_k = \frac{N_k}{\min_j(N_j)}, \quad j, k \in \{1, \dots, K\} \quad (1)$$

As mentioned earlier, K denotes the number of classes. We observe that the amount of imbalance varies across datasets, and the majority and minority classes may also differ when conditioned on the dataset. We now describe how we handle this imbalance for the FER task.

The goal of UDA is to perform well on the annotated source data and adapt to the target data simultaneously. Many UDA approaches [11, 27, 34] utilize two losses for this task, namely (i) classification loss on the annotated source samples and (ii) adaptation loss to bridge the domain gap between the source and target. Usually, the supervision from the labelled source domain \mathcal{D}_s is used to learn the final classifier by minimizing the Cross-Entropy(CE) loss. Specifically, given a sample $(x_s, y_s) \in \mathcal{D}_s$, let the corresponding output logits from the model be denoted as $\mathbf{z}_s = [z_1, z_2, \dots, z_K]^\top$. The CE loss is then computed as

$$\mathcal{L}_{CE}(\mathbf{z}_s, y_s) = -\log \left(\frac{\exp(z_{y_s})}{\sum_{k=1}^K \exp(z_k)} \right) \quad (2)$$

In general, if the training data is highly imbalanced, the standard cross-entropy loss may lead to poor generalization for minority classes because of the smaller margin between these classes. This problem due to data imbalance has been extensively explored in the supervised classification setting [4, 6, 14]. Here, we adopt the very successful Label Distribution Aware Margin (LDAM) approach [4] and modify it suitably for our task. The LDAM loss addresses the problem of class imbalance in classification by enforcing a class-dependent margin, with the margin being larger for the minority classes relative to that of the majority classes. Given the number of samples in each class $N_k, k \in \{1, \dots, K\}$, the class dependent margins $\Delta \in \mathcal{R}^K$ and the LDAM loss (termed here as Data Imbalance or DI loss) for a sample (x_s, y_s) is computed as

$$\mathcal{L}_{DI}(\mathbf{z}_s, y_s; \Delta) = -\log \frac{e^{z_{y_s} - \Delta_{y_s}}}{e^{z_{y_s} - \Delta_{y_s}} + \sum_{k \neq y_s} e^{z_k}} \quad (3)$$

where $\Delta_k = \frac{\gamma}{N_k^{1/4}}$ for $k \in \{1, \dots, K\}$

where γ is a hyperparameter. In [4], this loss, along with a deferred re-weighting scheme complementing each other, is proven to address the imbalance issue effectively for long-tail label distributed classification tasks without trading off the performance on the majority classes. Following this scheme, we weight this loss using class-specific weights after a few epochs. We use the subscript DI to

indicate that this is used for mitigating the effect of source data imbalance in the proposed DIFC module. Although this is very successful in handling data imbalance for supervised classification tasks, our goal is to correctly classify the target data. Fig. 1 shows that amount of imbalance varies across datasets, suggesting that the margins computed using the source label distribution as in eq. (3) may not be optimal for the target data. We now discuss how we address the target confusion along with the source data imbalance seamlessly in the proposed DIFC module.

4.2 Handling Feature Confusion in DIFC

The final goal is to improve the classification accuracy of the target features. Usually, samples from the minority classes tend to be confused with its neighbouring classes in the feature space. The target features can be confused because of the source as well as target data imbalance. But the target data imbalance may not be identical to that of the source. In addition, there may be other factors, e.g., few expressions are inherently close to one another compared to the others, which may result in increased confusion between certain classes. These unknown, difficult to quantify factors result in target feature confusion, which cannot be handled by LDAM or by aligning source and target distributions using UDA techniques. We now explain how this challenge is addressed in the proposed DIFC module.

The goal is to quantitatively measure and reduce the confusion between the target classes at any stage of the adaptation process. As the target data is unlabeled, one commonly used technique is to minimize the entropy of the unlabeled target data to enforce confident predictions, which in turn reduces the confusion among target classes. For this task, we empirically found that this decreased the target accuracy. Upon further analysis, we observed that several target examples were being pushed towards the wrong classes, which resulted in this performance decrease. This might be because classifying subtle facial expressions is, in general, very challenging and subjective. Thus, in the DIFC module, we try to find the confusion in a class-wise manner, hence reducing the influence of wrong class predictions of the individual target samples.

Here, we propose to use the softmax scores $\mathbf{p}_t = \text{softmax}(\mathbf{z}_t)$ predicted by the model to estimate a confusion score for each class. We first group the target samples $x_t \in \mathcal{D}_t$ based on the predictions made by the model into sets $S_k, k \in \{1, \dots, K\}$, where S_k contains all the target features classified into class k by the current model, i.e.,

$$S_k = \{x_t | \hat{y}_t = k\}; \quad \text{where} \quad \hat{y}_t = \text{argmax}(\mathbf{p}_t) \quad (4)$$

Using these predicted labels \hat{y}_t , we aim to quantify the confusion of each class in the target dataset. In general, samples from a particular class are confused with its neighbouring classes. But there are few classes which are more confusing than the others. Conditioned on the target dataset, this trend of confusion can differ from that of source. For a sample $x_t \in S_k$, if p_t^1 and p_t^2 are the top two

softmax scores in the prediction vector $\mathbf{p}_t \in \mathcal{R}^K$, we formulate the confusion score of class k as

$$\text{conf}_k = \frac{1}{\mathbf{E}_{x_t \in S_k} |p_t^1 - p_t^2|} \quad (5)$$

As $x_t \in S_k$, p_t^1 corresponds to its confidence score for class k , however the second score p_t^2 can correspond to any class $j \neq k$. This implies that, for all the target samples predicted to belong to class k , we compute the mean differences between the highest and second-highest softmax scores. If the mean difference is large, it implies that the model is not confusing its predicted class k with any other class, and the computed confusion score for class k is less. On the other hand, if for majority of the samples predicted to belong to a class, the average difference between the top-2 scores is small, it implies that this class is more confusing and thus gets a high confusion score. We emphasize that this confusion is not computed in a sample-wise manner as it can be noisy due to wrong predictions.

We propose to modify the margins in eq.(?) based on these estimated class confusion scores. Towards this goal, the class indices are sorted in decreasing order of confusion to get the ordered class indices C , such that $C[1]$ corresponds to the most confusing class and $C[K]$ corresponds to the least confusing class. We propose to use exponentially decreasing margin updates that emphasize on increasing the margin for the top few confusing classes without affecting the decision boundaries of the other classes. If $\Delta_{C[j]}$ denotes the margin for class $C[j]$, the class margins and DIFC loss are computed as follows:

$$\begin{aligned} \Delta'_{C[j]} &= \Delta_{C[j]} + \frac{\epsilon}{2^{j-1}} \\ \mathcal{L}_{DIFC} &= \mathcal{L}_{DI}(x_t, y_t; \Delta') \end{aligned} \quad (6)$$

where ϵ is a hyperparameter and refers to the increase in margin for the most confusing class i.e., $C[1]$. Using this formulation, we not only learn good decision boundaries taking into account the imbalanced source samples, but also reduce the confusion among the target domain samples.

4.3 Choosing baseline UDA approach

There has been a plethora of work in UDA for image classification, and it is not clear which one is better suited for this task. To choose a baseline UDA approach, we perform an empirical study of a few UDA approaches on SFEW2.0 [7] dataset, using RAF-DB [21] as the source data. The five approaches chosen are: (i) DANN [11]: classical UDA technique which forms the backbone for SOTA methods like AGRA [5], (ii) CADA [23]: approach which utilizes classifier outputs in a DANN-like framework, (iii) MCD [27]: approach which uses multiple classifiers to perform adversarial domain alignment, (iv) STAR [24]: very recent approach which extends the concept of multiple classifiers in MCD using a stochastic classifier, (v) SAFN [31]: a non-adversarial UDA approach that encourages learning features with larger norm as they are more transferable.

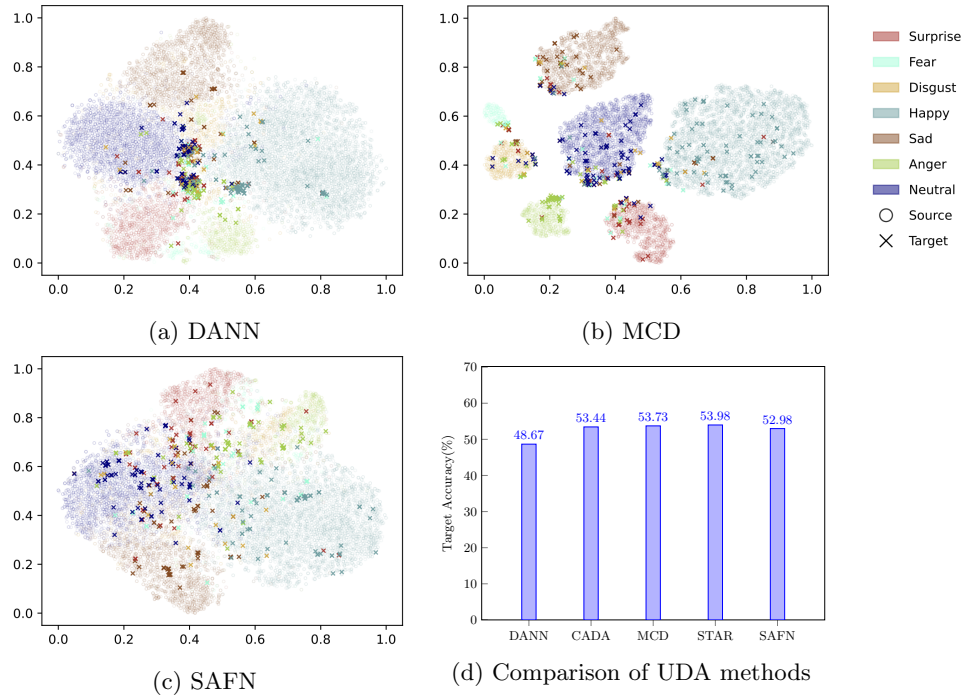


Fig. 2: We observe from the t-SNE plots for DANN(a), MCD(b) and SAFN(c) that the classes are better clustered and separated in MCD. (d) empirically shows the effectiveness of MCD when compared to others.

We observe from Fig. 2d that CADA [23], MCD [27] and STAR [24] achieve very similar accuracy, and all of them perform significantly better than DANN and SAFN. This implies that incorporating class information or class discrimination improves the classification performance of the target features. CD-FER task is characterized by large intra-class variations, small inter-class variations and domain shift. In such a fine-grained classification task, class-agnostic alignment based on domain confusion may not be sufficient. Confusing target samples often lie between two source clusters and could get aligned to the wrong class in this case. The effectiveness of using class discrimination is also evident from the t-SNE plots using DANN, SAFN and MCD (similar to CADA and STAR) in Fig. 2. Finally, in this work, we decided to use MCD [27] as the baseline UDA approach for the two following reasons: 1. It performs similar to the other approaches like CADA [23], STAR [24]. 2. Several recent works like SWD [17], MDD [34], STAR [24] build upon the principle of MCD. However, we emphasize that the proposed DIFC module can be seamlessly integrated with other UDA methods (based on completely different techniques) as an alternative for CE loss, which we demonstrate in Section 5.4.

Algorithm 1: MCD with DIFC module

Input:Labelled source domain data $\mathcal{D}_s = \{(x_s^i, y_s^i), i = 1, \dots, n_s\}$ Unlabelled target domain data $\mathcal{D}_t = \{x_t^i, i = 1, \dots, n_t\}$.Feature extractor F Two randomly initialized classifiers C_1 and C_2 .**Training:***Handling Data Imbalance*

Use Data Imbalance (DI) loss as classification loss.

for $i=1$ to T_1 epochsi.e $\mathcal{L}_{cls} \leftarrow \mathcal{L}_{DI}$ in eq. (??)Train F, C_1, C_2 using eq. (7).*Handling Feature Confusion*

Update margins and then use DIFC loss as classification loss.

for $i=T_1 + 1$ to T_2 epochsi.e. $\mathcal{L}_{cls} \leftarrow \mathcal{L}_{DIFC}$ in eq. (6)Train F, C_1, C_2 using eq. (7).**Output:**Trained feature extractor F and classifiers C_1 and C_2 .

4.4 Integrating DIFC with baseline UDA

Here, we briefly describe Maximum Classifier Discrepancy (MCD) [27] and how the DIFC module is integrated with it. MCD consists of three modules, a feature extractor F and two classifiers C_1 and C_2 . Here, the feature extractor acts as the generator, and the classifiers C_1 and C_2 play the role of the discriminator. The generator and classifiers are trained to learn accurate decision boundaries by minimizing cross entropy loss on source domain samples. In order to align the source and target domain features, the classifiers C_1 and C_2 are leveraged to identify target samples beyond the support of source by maximizing the prediction discrepancy between the classifiers. The generator F is then trained to minimize the discrepancy so that the target features align with that of source. Given $(x_s, y_s) \in \mathcal{D}_s$ and $x_t \in \mathcal{D}_t$, the generator F and classifiers C_1, C_2 are optimized in three steps as follows

$$\begin{aligned}
& \min_{F, C_1, C_2} \mathcal{L}_{cls}(C_1(F(x_s)), y_s) + \mathcal{L}_{cls}(C_2(F(x_s)), y_s) \\
& \max_{C_1, C_2} \|C_1(F(x_t)) - C_2(F(x_t))\|_1 \\
& \min_F \|C_1(F(x_t)) - C_2(F(x_t))\|_1
\end{aligned} \tag{7}$$

where \mathcal{L}_{cls} refers to the standard CE loss in MCD.

First, we propose to address data imbalance using \mathcal{L}_{DI} as the classification loss. Once the source samples are well separated, feature confusion is addressed

using the final DIFC module. As two classifiers are used in MCD, the class prediction is based on the average of the two scores given by

$$\begin{aligned}\mathbf{p}_{t,i} &= \text{softmax}(C_i(F(x_t))), \quad i = 1, 2 \\ \mathbf{p}_t &= \frac{1}{2}(\mathbf{p}_{t,1} + \mathbf{p}_{t,2}) \\ \hat{y}_t &= \text{argmax}_k(\mathbf{p}_t)\end{aligned}\tag{8}$$

The DIFC loss is then computed as described in eq. (6). During testing, given a target sample x_t , the class prediction is made using eq. (8). The final integrated algorithm (MCD with DIFC module) is described in Algorithm 1.

5 Experiments

5.1 Datasets Used

We demonstrate the effectiveness of the proposed method on several publicly available datasets spanning different cultures, pose, illumination, occlusions, etc. For fair comparison, we follow the benchmark [5] and use RAF-DB [21] as the source dataset. We use other FER datasets, namely JAFFE [25], SFEW2.0 [7], FER2013 [13] and ExpW [35] as the target data. For all the datasets, the images are labelled with one of the seven expressions, namely surprise, fear, disgust, happy, anger, sad and neutral.

5.2 Implementation Details

Model Architecture We use the same backbone as that used in [5] for fair comparison. A ResNet-50 backbone pretrained on MS-Celeb-1M dataset is used to extract local and global features. Firstly, MT-CNN [33] network is used to obtain the face bounding boxes as well as five landmarks corresponding to the two eyes, the nose and the two ends of the lips. Following [5], the face crop is resized to 112x112x3 image resolution and fed to the ResNet-50 backbone, from which a feature map of size 7x7x512 is obtained at the end of the fourth ResNet block. The global features are obtained by convolving this with a 7x7x64 filter followed by Global Average Pooling, resulting in a 64-dimensional feature vector. The feature maps from the second ResNet block of dimension 28x28x128 are used to extract the local features. This feature map is used to obtain five crops of 7x7x128 surrounding each corresponding landmark location. Each such feature map is further convolved with a 7x7x64 filter followed by Global Average Pooling, resulting in a 64-dimensional local feature descriptor for each landmark. The global and local features are then concatenated to obtain a 384 (64x(1+5)) dimensional feature. An additional fully connected layer enables global and local feature connections to get the final feature vector of 384 dimensions. We use two classifiers with different random initializations for MCD, each being a fully connected layer that does a 7-way classification of the 384-dimensional feature vectors.

Method	JAFFE	SFEW2.0	FER2013	ExpW	Avg
CADA [23]	52.11	53.44	57.61	63.15	56.58
SAFN [31]	61.03	52.98	55.64	64.91	58.64
SWD [17]	54.93	52.06	55.84	68.35	57.79
LPL [22]	53.05	48.85	55.89	66.90	56.17
DETN [19]	55.89	49.40	52.29	47.58	51.29
ECAN [20]	57.28	52.29	56.46	47.37	53.35
JUMBOT [9]	54.13	51.97	53.56	63.69	55.84
ETD [18]	51.19	52.77	50.41	67.82	55.55
AGRA [5]	61.50	56.43	58.95	68.50	61.34
Proposed DIFC	68.54	56.87	58.06	71.20	63.67

Table 1: Target classification accuracy (%) of the proposed framework compared with state-of-the-art methods.

Training details We train the backbone and the two classifiers using stochastic gradient descent (SGD) with an initial learning rate of 0.001, momentum of 0.9, and a weight decay of 0.0005. Initially, during the domain adaptation, only the DI loss is used for the initial epochs which addresses the source imbalance. The complete DIFC loss is used after few epochs to ensure that the target class predictions are more reliable and can be used to compute the feature confusion. In all our experiments, after training the model with only the DI loss for the initial 20 epochs, the learning rate is reduced to 0.0001 and the model is further fine-tuned with the complete DIFC loss for another 20 epochs. We tune γ to normalise the class margins $\Delta_j, j \in \{1, \dots, K\}$, so that the largest enforced margin is 0.3. We set ϵ to 0.02 for all the experiments. The same protocol without any modification is followed for all the target datasets. We use Pytorch framework and perform all experiments on a single NVIDIA GTX 2080Ti GPU using a batch size of 32.

5.3 Results on Benchmark Datasets

We report the results of the proposed framework on four datasets, i.e. JAFFE [25], SFEW2.0 [7], FER2013 [13] and ExpW [35] as the target and RAF-DB as the source in Table 1. Comparison with the state-of-the-art approaches is also provided. The results of the other approaches are taken directly from [5]. We observe that the proposed framework performs significantly better than the state-of-the-art for almost all datasets, except for SFEW2.0, where it is second to [5]. But on an average, it outperforms all the other approaches.

5.4 Additional Analysis

Here, we perform additional experiments to study the effect of different components of the loss function, integrate it with existing UDA methods. These show that our module is effective in widely varying experimental settings.

Method	JAFfE	SFEW2.0	FER2013	ExpW	Avg
CE	64.79	53.73	55.80	69.10	60.85
DI	67.14	55.90	57.30	70.70	62.76
DIFC	68.54	56.87	58.06	71.20	63.67

Table 2: Importance of both components of the proposed DIFC loss.

Dataset	Surprise	Fear	Disgust	Happy	Sad	Anger	Neutral
JAFfE	5	1	2	7	3	6	4
SFEW2.0	4	2	5	7	3	6	1
FER2013	6	2	1	4	3	7	5
ExpW	3	1	2	7	4	5	6

Table 3: Order of confusion for different target datasets

Ablation Study: To understand the impact of different components of the proposed DIFC loss, we perform experiments using MCD as baseline UDA method with CE (eq. 2) loss, DI or LDAM loss (eq. 3) and the proposed DIFC loss (eq. 6). The results in Table 2 show that using the DI loss improves target accuracy when compared to using the standard CE loss. Specifically, addressing source data imbalance improves the target accuracy by about 2% on average across all four target datasets. However, as the imbalance exists not only in the source but also in the target datasets, which may be different as shown in Fig. 1, the margins derived using source label distribution may not be optimal for the target dataset. The DIFC module adapts these margins based on the target feature confusion, further improving the target accuracy.

Analysis of confusion scores: In order to analyze the metric proposed to measure confusion (eq. 5), we compute the order of confusion obtained for each dataset after addressing the source data imbalance in Table 3. We observe that the confusing classes vary across datasets and also differ from the minority classes (Fig. 1). Based on this order, we divide the classes into two sets, *confusing* (with order 1,2,3) and *non-confusing* (with order 4,5,6,7). Since the total accuracy as reported in Table 1 can be biased towards the majority classes, here we analyze the average class accuracy for each of these two sets in Table 4 after handling data imbalance and further target confusion, which we refer as **DI** and **DIFC** respectively. We observe that (1) the average accuracy of the non-confusing classes is, in general, significantly more than that of the confusing classes, except for SFEW2.0, where they are very similar; (2) the DIFC module significantly improves the average accuracy for confusing classes without compromising the performance on the other classes. Table 6 shows a few model predictions from different target datasets for DI loss and the final DIFC loss. These results show that incorporating DIFC loss can correct several samples which were misclassified even after addressing source imbalance.

Dataset	Confusing classes		Non Confusing classes	
	DI	DIFC	DI	DIFC
JAFFE	48.01	51.22	80.88	80.88
SFEW2.0	50.70	51.89	46.88	48.02
FER2013	34.57	38.09	67.58	67.41
ExpW	30.78	31.20	66.25	66.75

Table 4: Average accuracy for confusing vs other classes.

Method	DANN	SAFN	MCD	Avg
CE	48.67	50.46	52.75	50.63
DI	50.12	51.81	53.25	51.73
DIFC	51.33	53.01	55.66	53.33

Table 5: DIFC with baseline UDA methods

Backbone: ResNet-18, Source: RAF-DB, Target: SFEW2.0

DIFC module integrated with other UDA approaches: Most UDA methods [11, 27, 31, 5] use CE loss to learn the classifier from the labelled source dataset and an adaptation loss driving the alignment between source and target features. However, the decision boundaries learnt using CE loss are not very effective in the presence of data imbalance. Here, we show that the DIFC module can be integrated with other UDA methods as an alternative for CE loss. We select three UDA approaches whose adaptation mechanisms are principally different from each other. DANN [11] and MCD [27] are adversarial methods that use domain discriminator and multiple classifiers respectively. As discussed, MCD incorporates class-discriminative information in the adaptation process unlike DANN. On the other hand, SAFN [31] is a non-adversarial UDA method. We use ResNet-18 backbone, RAF-DB as source and SFEW2.0 as target dataset in these experiments. Incorporating DIFC with each of these improves the target accuracy on SFEW2.0 dataset by about 2.5% when compared to using CE loss, as shown in Table 5.

Complexity Analysis The base feature extractor being common for AGRA [5] and our method, we report the additional number of parameters in the two methods. AGRA has a total of 318,464 extra parameters accounting for two intra-domain GCNs (2x64x18), inter-domain GCN (64x64), Classifier (384x7) and Domain discriminator (2x384x384+384x1). On the other hand, the proposed framework has only 152,832 additional parameters which is due to an extra FC layer (384x384) and two classifiers (384x7) used in MCD. The proposed method outperforms current state-of-the-art results using only about half the additional parameters when compared to AGRA.





























	Surprise	Fear	Disgust	Happy	Sad	Anger	Neutral
JAFPE							
	✓✓	X✓	X✓	✓✓	X✓	XX	✓✓
SFEW2.0							
	X✓	X✓	XX	✓✓	✓X	X✓	X✓
FER2013							
	✓✓	X✓	X✓	✓✓	✓✓	X✓	✓✓
ExpW							
	X✓	✓✓	X✓	✓X	✓✓	X✓	X✓

Table 6: Comparison of model predictions for DI (blue) and DIFC (orange) loss. The correct and incorrect predictions are marked with ✓ and ✗ respectively.

6 Conclusion

In this work, we propose a novel Data Imbalance and Feature Confusion (DIFC) module for the Cross-Dataset FER task. Firstly, the proposed module effectively mitigates the effect of source data imbalance and hence learns better decision boundaries. But this can be insufficient due to the shift in label distribution of the target data compared to the source data. To handle this and other unknown subjective factors that might be present, we devise the DIFC module to mitigate such confusion among the target classes. Specifically, the proposed DIFC module incorporates confusion in target data into the supervised classification loss of the baseline UDA framework to learn improved decision boundaries. Extensive experiments in varied settings demonstrate the effectiveness of the DIFC module for the CD-FER task. Additionally, the DIFC module can be seamlessly integrated with several existing UDA methods as an alternative for the standard CE loss, thereby further improving their performance.

Acknowledgements This work is partly supported through a research grant from AISIN.

References

1. Al-Modwahi, A.A.M., Sebetela, O., Batleng, L.N., Parhizkar, B., Lashkari, A.H.: Facial expression recognition intelligent security system for real time surveillance. In: Proc. of World Congress in Computer Science, Computer Engineering, and Applied Computing (2012)
2. Barandela, R., Rangel, E., Sánchez, J., Ferri, F.: Restricted decontamination for the imbalanced training sample problem. vol. 2905, pp. 424–431 (11 2003)
3. Brooks, J.A., Chikazoe, J., Sadato, N., Freeman, J.B.: The neural representation of facial-emotion categories reflects conceptual structure. *Proceedings of the National Academy of Sciences* **116**(32), 15861–15870 (2019)
4. Cao, K., Wei, C., Gaidon, A., Arechiga, N., Ma, T.: Learning imbalanced datasets with label-distribution-aware margin loss. In: *NeurIPS*. vol. 32 (2019)
5. 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* (2021)
6. Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S.: Class-balanced loss based on effective number of samples. In: *CVPR*. pp. 9268–9277 (2019)
7. Dhall, A., Goecke, R., Lucey, S., Gedeon, T.: Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In: *ICCV Workshops*. pp. 2106–2112 (2011)
8. Edwards, J., Jackson, H., Pattison, P.: Erratum to “emotion recognition via facial expression and affective prosody in schizophrenia: A methodological review” [clinical psychology review 22 (2002) 789–832]. *Clinical Psychology Review - CLIN PSYCHOL REV* **22**, 1267–1285 (11 2002)
9. Fatras, K., Sejourne, T., Flamary, R., Courty, N.: Unbalanced minibatch optimal transport; applications to domain adaptation. In: Meila, M., Zhang, T. (eds.) *ICML*. pp. 3186–3197 (2021)
10. Fragopanagos, N., Taylor, J.: Emotion recognition in human–computer interaction. *Neural Networks* **18**(4), 389–405 (2005), emotion and Brain
11. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., March, M., Lempitsky, V.: Domain-adversarial training of neural networks. *Journal of Machine Learning Research* **17**(59), 1–35 (2016)
12. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: *NeurIPS*. vol. 27 (2014)
13. Goodfellow, I.J., Erhan, D., Luc Carrier, P., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.H., Zhou, Y., Ramaiah, C., Feng, F., Li, R., Wang, X., Athanasakis, D., Shawe-Taylor, J., Milakov, M., Park, J., Ionescu, R., Popescu, M., Grozea, C., Bergstra, J., Xie, J., Romaszko, L., Xu, B., Chuang, Z., Bengio, Y.: Challenges in representation learning: A report on three machine learning contests. *Neural Networks* **64**, 59–63 (2015), special Issue on “Deep Learning of Representations”
14. Hayat, M., Khan, S., Zamir, S.W., Shen, J., Shao, L.: Gaussian affinity for max-margin class imbalanced learning. In: *ICCV* (October 2019)
15. Hoffman, J., Tzeng, E., Park, T., Zhu, J.Y., Isola, P., Saenko, K., Efros, A., Darrell, T.: CyCADA: Cycle-consistent adversarial domain adaptation. In: *ICML*. vol. 80, pp. 1989–1998 (2018)
16. Jack, R.E., Garrod, O.G.B., Yu, H., Caldara, R., Schyns, P.G.: Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences* **109**(19), 7241–7244 (2012)

17. Lee, C.Y., Batra, T., Baig, M.H., Ulbricht, D.: Sliced wasserstein discrepancy for unsupervised domain adaptation. In: CVPR. pp. 10277–10287 (2019)
18. Li, M., Zhai, Y.M., Luo, Y.W., Ge, P.F., Ren, C.X.: Enhanced transport distance for unsupervised domain adaptation. In: CVPR (June 2020)
19. Li, S., Deng, W.: Deep emotion transfer network for cross-database facial expression recognition. In: ICPR. pp. 3092–3099 (2018)
20. Li, S., Deng, W.: A deeper look at facial expression dataset bias. *IEEE Transactions on Affective Computing* pp. 1–1 (2020)
21. Li, S., Deng, W., Du, J.: Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: CVPR. pp. 2584–2593. IEEE (2017)
22. Li, S., Deng, W., Du, J.: Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: CVPR. pp. 2584–2593 (2017)
23. Long, M., CAO, Z., Wang, J., Jordan, M.I.: Conditional adversarial domain adaptation. In: NeurIPS. vol. 31 (2018)
24. Lu, Z., Yang, Y., Zhu, X., Liu, C., Song, Y.Z., Xiang, T.: Stochastic classifiers for unsupervised domain adaptation. In: CVPR. pp. 9108–9117 (2020)
25. Lyons, M., Akamatsu, S., Kamachi, M., Gyoba, J.: Coding facial expressions with gabor wavelets. In: *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*. pp. 200–205 (1998)
26. Murez, Z., Kolouri, S., Kriegman, D., Ramamoorthi, R., Kim, K.: Image to image translation for domain adaptation. In: CVPR. pp. 4500–4509 (2018)
27. Saito, K., Watanabe, K., Ushiku, Y., Harada, T.: Maximum classifier discrepancy for unsupervised domain adaptation. In: CVPR (2018)
28. Sajjad, M., Zahir, S., Ullah, A., Akhtar, Z., Muhammad, K.: Human behavior understanding in big multimedia data using cnn based facial expression recognition. *Mobile networks and applications* **25**(4), 1611–1621 (2020)
29. Wang, X., Wang, X., Ni, Y.: Unsupervised domain adaptation for facial expression recognition using generative adversarial networks. *Computational intelligence and neuroscience* (2018)
30. 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* (2020)
31. Xu, R., Li, G., Yang, J., Lin, L.: Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. In: ICCV (October 2019)
32. Yan, K., Zheng, W., Cui, Z., Zong, Y.: Cross-database facial expression recognition via unsupervised domain adaptive dictionary learning. In: NeurIPS. pp. 427–434 (2016)
33. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters* **23**(10), 1499–1503 (2016)
34. Zhang, Y., Liu, T., Long, M., Jordan, M.: Bridging theory and algorithm for domain adaptation. In: ICML. pp. 7404–7413 (2019)
35. Zhang, Z., Luo, P., Loy, C.C., Tang, X.: Learning social relation traits from face images. In: ICCV. pp. 3631–3639 (2015)
36. Zheng, W., Zong, Y., Zhou, X., Xin, M.: Cross-domain color facial expression recognition using transductive transfer subspace learning. *IEEE transactions on Affective Computing* **9**(1), 21–37 (2016)
37. Zhou, L., Fan, X., Ma, Y., Tjahjedi, T., Ye, Q.: Uncertainty-aware cross-dataset facial expression recognition via regularized conditional alignment. In: *Proceedings of the 28th ACM International Conference on Multimedia*. p. 2964–2972 (2020)