The real-world facial expression recognition (FER) datasets suffer from noisy annotations due to crowd-sourcing, ambiguity in expressions, the subjectivity of annotators and inter-class similarity. However, the recent deep networks have strong capacity to memorize the noisy annotations leading to corrupted feature embedding and poor generalization. To handle noisy annotations, we propose a dynamic FER learning framework (DNFER) in which clean samples are selected based on dynamic class specific threshold during training. Specifically, DNFER is based on supervised training using selected clean samples and unsupervised consistent training using all the samples. During training, the mean posterior class probabilities of each mini-batch is used as dynamic class-specific threshold to select the clean samples for supervised training. This threshold is independent of noise rate and does not need any clean data unlike other methods. In addition, to learn from all samples, the posterior distributions between weakly-augmented image and strongly-augmented image are aligned using an unsupervised consistency loss. We demonstrate the robustness of DNFER on both synthetic as well as on real noisy annotated FER datasets like RAFDB, FERPlus, SFEW and AffectNet. Our source codes are publicly available at https://github.com/1980x/DNFER.

Keywords Facial expression recognition, Noisy annotations, Dynamic training, Consistency, Strong augmentation, Weak-augmentation

1 Introduction

Facial expressions recognition (FER) is aimed at making machines understand human emotions and intentions. In the recent years, a significant progress is made towards developing deep learning (DL) based robust FER systems [1, 2, 3, 4, 5, 6] along with development of large-scale datasets like RAFDB [7, 8], FERPlus [9], AffectNet [10], SFEW [11] etc. However, large real-world datasets suffer from noisy annotations due to subjectivity of annotators, ambiguity of different expressions, poor quality of images etc. A sample of images from RAFDB, FERPlus and AffectNet are shown in Fig. 1. Clearly, some of them have reliable labels whereas others have uncertain labels. These uncertain labels lead to noisy annotations leading to corrupted features and poor generalization as shown in the Fig. 2. Here, we train ResNet-18 [12] on RAFDB with standard cross-entropy loss (denoted as Baseline) in the presence of 30% synthetic label noise and compare the performance on train and test sets during training. As training progress, Baseline model overfits noisy samples (shown in solid brown color) but its performance on test set (shown in dotted brown color) is
much lower compared to the proposed method (shown in blue color). This is due to strong memorization capability of deep networks [13, 14], as network overfits on noisy annotated samples.

![Figure 1: Examples of samples with certain and uncertain annotations](image)

Figure 1: Examples of samples with certain and uncertain annotations

![Figure 2: Performance comparison of train and test set accuracy on RAFDB with 30% synthetic label noise with Baseline (supervised training) and DNFER (proposed)](image)

Figure 2: Performance comparison of train and test set accuracy on RAFDB with 30% synthetic label noise with Baseline (supervised training) and DNFER (proposed)

Recently, some works like [15, 16, 17, 18, 19, 20] have studied the problem of uncertain labels in FER but their performance is not satisfactory. Self-cure network (SCN) [15] attempts to eliminate the influence of noisy annotated samples by loss re-weighing and relabelling incorrect ones based on learnt importance scores but it suffers from self-confirmation bias. ULC-AG [17] improvises over SCN by using auxiliary action unit (AU) graphs for uncertain label correction. LDL [18] learns the label distribution whereas DMUE mines the latent distribution and uses pairwise uncertainty estimation for ambiguity estimation. In [16], likelihood of sample being clean is obtained based on Jensen Shannon Divergence between prediction probabilities and soft ground-truth labels. Noise robust loss function called Rayleigh loss [21] is proposed to learn discriminative expression features in the presence of noisy annotations. However, many of these methods fail to perform well and most of these are tested only on synthetic noisy datasets. Robust training under noisy labels is an important problem in DL. Recently, many approaches have been developed like robust loss functions [22, 23], loss correction [24, 25] and sample selection methods [26, 27, 28, 29, 30], etc. Sample selection methods like Co-teaching [27], Co-teaching+ [28], JoCoR [29] and NCT [30] train two networks with peer learning or joint learning using selected clean samples based on small loss criterion but they need to know the noise rate to select small loss samples. Further, threshold chosen to select clean samples is same for all the classes and fixed throughout of training which may not be true. This is demonstrated in the Fig. 3 (a). The average posterior prediction probabilities of different expression classes Happy, Neutral, Sad, Fear, Disgust, Surprise, Anger and Contempt for benchmark FER datasets RAFDB (except contempt), FERPlus and AffectNet are plotted during the first 10 epochs of supervised training. These average class probabilities are different for each class as well as vary while training. In addition, the number of samples are imbalanced in terms of different expression classes for different datasets as shown in Fig. 3 (b). So, fixing a same prior threshold for all classes is likely to select samples only from majority classes which leads to poor performance on other minority classes. Further, having a same threshold for each mini-batch throughout training prevents selection of hard samples. Motivated by this observation, we propose a dynamic adaptive threshold for selecting clean (reliable) samples based on average posterior prediction probabilities of samples in each mini-batch. The selected clean samples are trained using supervision loss based on cross-entropy (CE) loss. Further, in order to make use of all the samples for feature learning, the posterior distributions of weakly-augmented image
and strongly-augmented image of each sample is minimized using unsupervised consistency loss based on symmetric K\text{L}-divergence loss. The proposed framework uses a single shared network for noise robust FER learning based on dynamic adaptive threshold named as DNFER. The proposed DNFER framework is illustrated in Fig. 4 and explained in Section 3. Effectiveness of the DNFER framework is validated on synthetic as well as real noisy annotated FER datasets.

Overall, the main contributions can be summarized as follows:

1. The proposed DNFER selects clean (reliable) samples for supervised learning based on dynamic class adaptive threshold and utilizes noisy (unreliable) samples using consistency loss for facial expression representation learning.

2. DNFER is architecture independent without depending upon noise rate as well as separate clean data. It uses a data-dependent class-specific dynamic threshold to select clean samples in each mini-batch.

3. Our DNFER is an end-to-end framework and achieves superior performance on several benchmark FER datasets as well as on synthetic and real noisy annotated datasets.

The rest of the paper is organized as follows: Section 2 reviews the related work in the area of FER in general and with noisy label learning as well as general DL based noisy robust methods. Section 3 presents the proposed DNFER framework. The effectiveness of DNFER is demonstrated by comparative performance evaluation and ablation study in Section 4. Finally, Section 5 concludes the work.

2 Related Work

In this section, we first briefly review the related work in FER with and without noisy labels followed by noisy label learning.

2.1 Facial Expression Recognition

Recent DL based FER methods are aimed at either representation learning [1, 2, 3, 4, 5, 6] or uncertain label based learning [31, 19, 18, 15, 17]. Attention based CNN methods like SCAN [2], OADN [1], GACNN [32], RAN [4] learn local and global features based on attention scores of different facial regions for learning robust expression features in the presence of occlusion and pose variations. FDRN [5] uses feature decomposition and reconstruction networks respectively for capturing similarities across different expressions and distinctive feature for each expression. Efficient-face [6] is a light-weight model designed using dept-wise convolution and channel modulator for capturing local-global features for robustness against occlusion and pose variations. In the same line, CERN [33] is a low-calorie power-packed architecture with 1.45M parameters which is robust to occlusion and pose variations. Although, DL based methods have achieved great success for robustness against occlusion and pose variations in FER, their unique and strong memorization learning ability makes them overfit on the noisy annotations [13, 14].
Recently, some works like SCN [15], ULC-AG [17], DMUE [19], LDL [18], IPA2LT [34], CCT [20] and RUL [31] have attempted to mine the latent label distribution to correct the uncertain labels in FER datasets. SCN [15] learns the confidence score of each sample for loss-reweighing using a self-attention module. It corrects the labels of uncertain samples (low score samples) based on prediction probability. In Inconsistent Pseudo Annotations to Latent Truth (IPA2LT), multiple inconsistent human and machine predicted labels are assigned to each sample, and subsequently the true label is learnt by maximizing the log-likelihood of these inconsistent annotations. Rayleigh loss [21] introduces a constraint to ensure that class center is far away from nearest neighbours along with center loss. RUL [31] utilises a model to learn uncertainties of input images and uses these uncertainties to mix the features of two different labels. In CCT [20], three networks are trained using a convex combination of supervision loss and consistency combined using a dynamic balancing scheme to combat the influence of noisy annotated samples. LDL [18] proposed a label distribution learning using auxiliary-label space graphs based on action-units similarity with different expressions. DMUE [19] designed an uncertainty estimation module which dynamically learns from mined label distribution and ground truth label of the samples. Though these methods mine the label distribution but these have not observed effect of noisy labels in the presence of inter-class sample size imbalance and intra-class imbalance of difficulty level. To mitigate this, a dynamic class adaptive learning based framework is proposed for robust expression representation learning in the presence of noisy annotated samples.

2.2 Noisy Labels Robust Deep Learning

Learning with noisy labels is an important research problem in DL. Recently, it has been observed that DNNs first fit clean data (data with correct labels), and then, towards later part of the training, memorize examples with noisy annotations [13, 14]. This makes it challenging to train models efficiently on real world datasets in which poor quality annotations are obtained using online queries or crowd sourcing. Based on above, many methods have been proposed which can be categorized as sample selection using small loss [26, 27, 28], label cleansing [35], weighting [36], robust loss functions [22, 23] etc. Please refer to survey [37] for detail on these methods. This paper adopts the strategy of sample selection, so we briefly review these here. These methods rely on selecting small loss samples with or without co-training [38, 28, 29, 27, 30] due to the fact that small loss samples are associated with clean labels. Inspired from semi-supervised learning consisting of labelled and unlabelled data, Co-training [39] principle has been followed in these works. Co-training assumes that each data has two different views and each view is sufficient for learning an effective model. These views can either correspond to different data sources or different representations, and provide complementary information about each sample. Co-training assumes that models trained on two different views have consistent predictions on unlabelled data. Co-training based methods follow the principle of joint training with consensus in JoCoR [139] or mutual training with consensus in NCT [30] or peer training with cross supervision in Co-teaching [27] and Co-teaching+ [28]. Co-teaching and JoCoR follow the principle of agreement whereas Decoupling and Co-teaching+ depend upon the “Disagreement” strategy to keep the two networks diverged to achieve better ensemble effects. To summarize, recent methods co-train two networks using small loss samples based on agreement or disagreement combined with some kind of regularization. These methods need to know noise rate to drop high loss samples in each mini-batch which may be difficult to estimate in real world setting. Further, threshold chosen based on noise rate to select clean samples is same for all the classes and fixed throughout training which may not be true. Recent works like CCT [20] and Co-curing [40] have applied these methods in FER scenario but these use multiple networks for training which leads to computational burden. Further, they use a fixed threshold for all expression classes but inherently FER datasets are imbalanced in sample size as well as intra-class samples have serious imbalance in the difficulty level. This imbalance factor has not been considered for selecting samples of small loss during learning.

3 Proposed Method

3.1 DNFER Overview

The proposed DNFER method is based on the idea of small loss (reliable) sample selection [13, 14] using a dynamic adaptive threshold and uses 2 different transformed (augmented) views of an image based on co-training [39] for learning noisy label robust expression feature representation and classification. Unlike the previous works [27, 28, 29], which use a fixed threshold for all classes for selection of small loss samples depending upon the noise rate in the data (which may not be known and otherwise difficult to estimate), our method dynamically learns the adaptive threshold based on the average posterior probabilities of each mini-batch without depending upon noise rate in the given data. Since posterior probabilities are inversely related to loss values in the categorical CE loss for supervision, so, small loss samples (high posterior probabilities) are likely to correspond to clean (reliable labels). DNFER uses a dynamic class adaptive threshold which improves along with learning taking into account intra-class sample difficulty as well as inter-class sample size imbalance. The selected clean samples are trained using CE loss. Further, unreliable samples
are used for expression feature representation by aligning the posterior distribution of weak-augmented image and strongly-augmented image using unsupervised consistency loss. The supervision loss and consistency loss are combined using a weighing parameter (chosen based on ablation study). The overview of the proposed DNFER framework is shown in Fig.4.

Figure 4: Proposed DNFER framework: a) Given an input face image \( x \) with expression label \( y \), its weak-augmented image \( (x_w) \) and strongly-augmented image \( (x_s) \) are input to a shared network \( f \) with parameter \( \theta \). Clean samples are selected based on dynamic class specific threshold \( (T_c) \) computed using average posterior probabilities of samples in each mini-batch. Model is trained using supervision loss based on selected clean (reliable) samples and unsupervised consistency loss by aligning the posterior predictions \( p_w \) and \( p_s \) of all samples.

### 3.2 Problem formulation

Given a batch of \( N \) images \( S = \{(x_i, y_i)\}_{i=1}^N \) where each facial image \( x_i \) has an expression label \( y_i \in \{1, 2, ..., C\} \) and \( C \) denotes the number of expression classes. The shared backbone network is parameterized by \( \theta \) (in this work ResNet-18 is used as the backbone). Features from backbone are classified using a fully connected layer (FC) layer with softmax to obtain prediction probabilities. The weak-augmented image and strongly-augmented image are denoted as \( x_w \) and \( x_s \) and their prediction probabilities as \( p_w \) and \( p_s \) respectively. Standard cropping along with horizontal flipping are used for weak augmentations. Randaugment [41] is used as strong augmentation. During training, the small loss samples from each mini-batch \( S_n \) are selected based on dynamic adaptive confidence threshold \( T_c \) (explained below). The model is trained by minimizing a loss \( L \) which is a convex combination of supervision loss (based on standard CE loss) denoted as \( L_{sup} \) and consistency loss (based on symmetric KL divergence) denoted as \( L_{cons} \). The overall loss is given as below:

\[
L = \alpha * L_{cons} + (1 - \alpha) * L_{sup}
\]

Here, \( \alpha \in [0, 1] \).

### 3.3 Dynamic Adaptive Threshold based Clean Sample Selection

As deep networks tend to fit easy (small loss) samples before the noisy (high loss) ones, we use CE loss for training the model using ground truth labels, denoted as \( L_{CE}(X, Y; \theta) \),

\[
L_{CE} = \{l_i\}_{i=1}^N = \{- \sum_{c=1}^{C} y_{i,c} \log(p_c^r(x_i; \theta))\}_{i=1}^N
\]

where \( x_i \in X, y_i \in Y \), \( p_c^r \) is observed posterior probability for class \( c \in \{1, 2, ..., C\} \) from network’s softmax layer. Clean samples are distinguished from noisy ones according to \( L_{CE} \). Since, the loss and posterior probabilities are inversely related, we use class wise mean of posterior probabilities of samples in a mini-batch to obtain class wise...
threshold \( T_c \). It assumes that all sample losses are independent and identically distributed. Samples with predicted probabilities greater than the dynamic threshold \( T_c \) are treated as clean samples. \( T_c \) is computed as:

\[
T_c = \frac{1}{|S_c|} \sum_{x \in S_c} p_c(x; \theta)
\]

(3)

where \( S_c \) is set of samples in the current mini-batch with class \( c \). This dynamic class wise adaptive threshold evolves along with learning depending upon the mini-batch and takes into account the intra-class difficulty of image from same expression class as well as inter-class imbalance of sample sizes for different expression classes. As shown in Fig 3, happiness, neutral, surprise and sad expression classes are easily recognizable compared to fear, disgust and contempt in real world FER datasets [7, 9, 10]. Due to which, these datasets are imbalanced in terms of sample size. Apart from this, within one expression class, some images are easy and others are hard apart from noisy annotated samples. As deep networks fit easy samples before hard and noisy, so network is trained using CE loss on easy samples using ground truth labels for warm up epochs (hyper-parameter based on ablation study) which results in different loss distribution for different expression classes. Average of class probabilities are different for different expression classes as shown in the Fig. 3 (a) and these change as training evolves. So, this dynamic class specific adaptive threshold adopts intra-class and inter-class imbalance for selecting clean samples. Finally, based on dynamic threshold \( T_c \), clean-sample selection is done as follows:

\[
S_{c}^{\text{clean}} = \{ x \in S_c \mid p(x; \theta) \geq T_c \}
\]

(4)

### 3.4 Supervision loss

After the warm-up epochs, network is trained based on supervision loss using selected clean samples. Inspired by Co-training methods [39, 27, 29] for semi-supervised learning and noisy label training, supervision loss is computed based on selected clean samples for two different transformed views of input samples in each mini-batch. The posterior probabilities of selected clean weak-augmented (\( x_{\text{clean}}^{w} \)) and strongly-augmented (\( x_{\text{clean}}^{s} \)) transformed images are denoted as \( p_{w}^{\text{clean}} \) and \( p_{s}^{\text{clean}} \) respectively.

\[
L_{\text{sup}} = L_{\text{CE}}(y^{\text{clean}}, p_{w}^{\text{clean}}) + L_{\text{CE}}(y^{\text{clean}}, p_{s}^{\text{clean}})
\]

(5)

where \( y^{\text{clean}} \) is the ground-truth labels of selected clean samples (\( S_{c}^{\text{clean}} \)) in current mini-batch.

### 3.5 Consistency loss

Finally, in order to utilize all the samples for learning, the posterior distributions of weak-augmented and strongly-augmented images are aligned using unsupervised consistency loss according to Eq. (6). The symmetric Kullback-Leibler divergence (\( D_{KL} \)) defined in Eq. (7) is used to compute consistency loss.

\[
L_{\text{cons}} = D_{KL}(p_{s}||p_{w}) + D_{KL}(p_{w}||p_{s})
\]

(6)

\[
D_{KL}(p_{k}||p_{j}) = \sum_{l=1}^{C} p_{k}^{l} \log \frac{p_{k}^{l}}{p_{j}^{l}}
\]

(7)

### 3.6 DNFER training algorithm

During the warm-up period, all the samples are used for training based on supervision loss. After warm-up, model is trained using \( L_{\text{sup}} \) and \( L_{\text{cons}} \). Details of training are provided in Algorithm 1. The loss weighing factor \( \alpha \) to combine \( L_{\text{sup}} \) and \( L_{\text{cons}} \) in Eq. (1) is fixed as 0.5 (here warm-up period as well as loss weighing parameter are chosen based on ablation study).

### 4 Experiments

The proposed DNFER is compared with recent state-of-the-art (SOTA) methods on popular real-world benchmarks as well as in the presence of synthetic label noise on FER datasets. Further, we conduct extensive experiments to verify the component wise contribution of each part of DNFER. Ablation studies are conducted for hyperparameter selection. Experimental results demonstrate that our DNFER is robust in presence of noisy annotations present in FER datasets. First, we briefly introduce the experimental setup and next present the performance evaluation followed by ablation studies and comparison with SOTA methods.
Algorithm 1: DNFER training algorithm

Input: Given a model \( f \) with parameters \( \theta \), dataset \( S = \{ (x_i, y_i) \}_{i=1}^{N} \), mini-batch size \( b \), learning rate \( \eta \), number of expression classes \( C \), total epochs \( E_{\text{max}} \), warm-up epochs \( E_{\text{warm}} \).

Output: Updated model parameters \( \theta \)

1. Initialize \( \theta \) randomly.
2. for \( e = 1, 2, ..., E_{\text{max}} \) do
   3. Shuffle training samples \( \{ (x_i, y_i) \}_{i=1}^{N} \)
   4. Sample mini-batch \( S_n \) from \( S \)
   5. for each class \( c \in \{1, 2, ..., C\} \) do
      6. Compute dynamic confidence threshold \( T_c \) using Eq. (3)
      7. Select clean samples \( S_n^{\text{clean}} \) from current mini-batch using Eq. (4)
   8. Compute supervision loss \( L_{\text{sup}} \) on above selected clean samples using Eq. (5)
   9. Compute consistency loss \( L_{\text{cons}} \) on all samples using Eq. (6)
   10. Compute total loss \( L \) using Eq. (1)
   11. if \( e < E_{\text{warm}} \) then
      12. \( \alpha = 0 \)
      13. else
      14. \( \alpha = 0.5 \)
      15. Update \( \theta = \theta - \eta \nabla L \theta \)
16. return \( \theta \)

4.1 Experimental Setup

4.1.1 Datasets

We evaluate DNFER on four commonly used datasets RAFDB, FERPlus, SFEW and AffectNet.

- **RAFDB** [7, 8]: The Real-world Affective Face Database (RAFDB) is constructed using 29762 face images with two different subsets: i) basic emotion subset annotated with 7 emotions (i.e. happy, sad, neutral, surprise, disgust, angry, fear), ii) compound emotion subset. In this study, we use basic emotion set of 12271 images for training and 3068 images for testing. Both train and test sets are imbalanced w.r.t sample size of different expression classes.

- **FERPlus** [9]: FERPlus is an extended version of FER2013 [42]. It consists of images with the 8-basic emotions (with contempt), of which 28709 are used for training, 3589 are used for validation, and the remaining 3589 for test. As 10 independent annotators are used for labeling each image, label quality of FERPlus is superior compared to FER2013. For a fair comparison, majority voting is used for reporting performance.

- **SFEW** [11]: SFEW contains images from movies with seven basic emotions with train set of 891 samples, validation set of 431 samples and test set of 372 samples. Performance is reported on the validation set as test sets labels are not released.

- **AffectNet** [10]: AffectNet is the largest dataset with 0.44M manually annotated and 0.46M automatically annotated facial expression images for 8 emotions. Main limitation is that each image has been labelled with a single annotator due to which labels are not reliable. AffectNet-7 refers to manually annotated set without contempt class whereas AffectNet-8 includes all expression images. AffectNet-7 has 283901 images for training and 3500 images for validation whereas AffectNet-8 has 287568 images for training and 4000 images for validation. Test set labels are not released, so, being consistent with previous works, validation sets are used for reporting results. In addition, we use *Automatically annotated subset* for training under real noisy conditions.

- **Synthetic noisy annotated datasets**: We randomly change 10%, 20% and 30% labels of training images from RAFDB, FERPlus and AffectNet-8 datasets to create synthetic noisy annotated datasets. Performance of DNFER is reported on the corresponding clean test/validation sets.
4.1.2 Performance Metrics

We use overall test accuracy (%) to report the performance for all algorithms. Beside we conduct experiments three times using different random seeds to obtain the mean accuracy and their standard deviations. Confusion plots are provided to obtain individual expression performance.

4.1.3 Implementation Details

In the below experiments, we use MTCNN [43] to detect and resize the facial expression images with size of 224x224. Our proposed DNFER framework is implemented in Pytorch DL toolbox using single GeForce GTX TitanX GPU with 12GB RAM. Backbone network used is ResNet-18 which is pretrained on MS-Celeb-1M [44] face recognition dataset similar to previous works. Random horizontal flipping with probability 0.5 is used for weak augmentation beside random crop with 4 pixel and resize to 224x224. For strong augmentation, RandAugment [41] is used. RandAugment randomly select two transformations from a group of transformations like contrast adjustment, rotation, color inversion, translation, etc. Oversampling\(^1\) is used to overcome class imbalance problem in AffectNet similar to [4, 15, 1]. During training, the batch size used is 128, the max epochs (E\(_{\text{max}}\)) is set as 40 for RAFDB, FERPlus datasets and 20 for AffectNet dataset, warm up period is chosen as 5 for RAFDB, FERPlus datasets and 3 for AffectNet dataset. Model is optimized using Adam optimizer with an initial learning rate of 0.001 which is decayed exponentially with a factor of 0.95 every epoch.

4.2 Experiment Results on Synthetic Noisy Annotated Datasets

Table 1: Performance evaluation (accuracy %) on FER datasets with synthetic label noise

| Noise-level | Method       | RAFDB    | FERPlus  | AffectNet-8 |
|-------------|--------------|----------|----------|-------------|
| 10%         | Baseline     | 81.74    | 85.87    | 57.21       |
|             | SCN [15]     | 82.18    | 84.28    | 58.58       |
|             | RR [21]      | 82.43    | 83.93    | 60.04       |
|             | ULC-AG [17]  | 83.21    | -        | 59.45       |
|             | DMUE [19]    | 83.19±0.83| -         | 61.21±0.36  |
|             | **DNFER**    | 88.09±0.37| 88.67±0.22| 59.94±0.08  |
| 20%         | Baseline     | 79.60    | 84.02    | 56.23       |
|             | SCN          | 80.01    | 83.17    | 57.25       |
|             | RR           | 80.41    | 83.55    | 58.47       |
|             | ULC-AG       | 81.16    | -        | 58.51       |
|             | DMUE         | 81.02±0.69| -         | 59.06±0.34  |
|             | **DNFER**    | 86.94±0.17| 87.95±0.26| 58.63±0.18  |
| 30%         | Baseline     | 74.31    | 82.30    | 52.60       |
|             | SCN          | 77.46    | 82.47    | 55.05       |
|             | RR           | 76.77    | 82.75    | 55.82       |
|             | ULC-AG       | 79.01    | -        | 56.45       |
|             | DMUE         | 79.41±0.74| -         | 56.88±0.56  |
|             | **DNFER**    | 84.99±0.35| 86.58±0.16| 56.55±0.32  |

Synthetic symmetric noise is manually added on RAFDB, FERPlus and AffectNet-8 datasets by randomly changing labels in the ratio of 10-30%. We compare DNFER with single ResNet-18 trained using CE loss (referred as Baseline) and against recent noise robust SOTA FER methods like SCN [15], RR [21], ULC-AG and DMUE. Results are presented in Table 1. Clearly, their performance is much inferior compared to our proposed DNFER in the presence of different noise levels on RAFDB and FERPlus (ULC-AG and DMUE don’t report on FERPlus). There is gain of around 5% for different noise levels on RAFDB compared to next best performing DMUE. DNFER performs best on FERPlus compared to Baseline as well as SCN and Rayleigh loss (denoted as RR). AffectNet-8 being a challenging noisy dataset because of heavy class imbalance as well as intra class difficulty, no method is able to achieve good performance. DNFER obtains comparable performance to that DMUE on AffectNet-8 in the presence of noisy labels.

\(^1\)https://github.com/ufoym/imbalanced-dataset-sampler
4.3 Evaluation on Real Noisy Annotated Dataset

Unlike synthetic label noise, real-noisy annotated FER datasets can include label noise of unknown level. We train DNFER using automatically annotated subset of AffectNet-7 and test on corresponding validation set. We compared the performance w.r.t recent SOTA FER methods like RAN, GACNN and OADN and noise robust FER methods like SCN and ULC-AG. As shown in Table 2, DNFER performs best in the presence of real uncertainties compared to general as well as noisy label robust FER methods ULC-AG and SCN. Uncertainties in real datasets are due to imbalance of sample size in different classes as well as imbalance of intra class sample at difficulty level. Due to adaptive threshold for different classes dynamically evolving along with learning, DNFER mitigates these real ambiguities effectively compared to recent methods.

Table 2: Performance evaluation on real noisy annotated AffectNet-7 (* denotes FER methods with noisy labels robust learning)

| Method     | Year | AffectNet-7 |
|------------|------|-------------|
| Baseline   | -    | 53.85       |
| RAN [4]    | 2019 | 56.43       |
| GACNN [32] | 2019 | 52.43       |
| OADN [1]   | 2020 | 55.37       |
| SCN* [15]  | 2020 | 55.43       |
| ULC-AG* [17]| 2022| 57.37       |
| DNFER      | -    | 57.65       |

4.4 Ablation Study

4.4.1 Evaluation of components

DNFER is aimed at the influence of noisy annotated samples for supervision learning based on selected clean samples using dynamic adaptive threshold and making use of all samples for feature representation learning using consistency loss. In this experiment, we design 4 settings for validating the effectiveness of the proposed DNFER. In the first setting, model is trained using all samples based on CE loss for all 40 epochs (referred to as Baseline). In the second setting, DNFER (only $L_{cons}$), model is trained with CE loss for warm-up epochs and later using only $L_{cons}$ i.e. without $L_{sup}$ loss on selected clean samples. In the third setting, DNFER (only $L_{sup}$), model is trained with CE loss for warm-up epochs and later with only $L_{sup}$ loss on selected clean samples without $L_{cons}$. Last one refers to our proposed DNFER. We evaluate the performance in the presence of 0-30% on RAFDB in the Fig. 5 and on FERPlus in the Fig. 6. Baseline performance on RAFDB (0% label noise) is inferior compared to DNFER with $L_{cons}$ and DNFER with $L_{sup}$ whereas it is not same on FERPlus (0% label noise). This indicates that FERPlus has cleaner annotations compared to RAFDB. As noise labels increase from 10-30% on RAFDB and FERPlus, both DNFER with $L_{cons}$ and DNFER with $L_{sup}$ fail to perform. DNFER using both the components obtains superior performance by selecting clean annotated samples for supervised training and effectively learn the expression features using all the samples in the presence of noisy annotations.

Figure 5: Evaluation of the components in DNFER on RAFDB for different noise levels of 0-30%
4.4.2 Impact of Weighting Parameter

We analyse the impact of loss weighing parameter $\alpha$ in Eq. (1) to combine the supervision and consistency loss. Fig. 7 depicts the performance against various value of $\alpha$ on RAFDB with 30% (shown in blue color) and 0% (shown in orange color) synthetic noise. Clearly, $\alpha = 0.5$ gives best performance.

![Figure 7: Impact of weighing parameter $\alpha$ on RAFDB (with and without noise)](image)

4.4.3 Influence of Warm-up Epochs

In order to analyse the influence of different number of warm-up epochs, we compare the performance on RAFDB with 30% (shown in orange color) and 0% (shown in blue color) synthetic noise as shown in the Fig. 8. As deep networks learn from easy (small loss) samples during initial stage of training but tend to overfit on noisy annotated samples as training progress, we choose warm-up epochs as 5 for reporting results on RAFDB and FERPlus datasets and 3 on AffectNet.

4.4.4 Comparison of noise over-fitting

In order to analyse the influence of DNFER in mitigating overfitting on noisy annotated samples, we plot the performance on train and test set in the presence of 30% synthetic label noise on RAFDB (shown in Fig. 2), FERPlus (shown in Fig. 9) and AffectNet (shown in Fig. 10) datasets. It is clear that using Baseline model based on all samples, training accuracy (shown in sold grey curve) saturates whereas test accuracy (shown in dotted curve grey) initially increases but then drops off as training progress. In contrast, DNFER training accuracy (shown in solid blue curve) is much lower compared to Baseline train accuracy and test accuracy curve (shown in dotted blue) is much above to that of Baseline test accuracy. This behaviour is observed across all the three datasets demonstrating the validity of DNFER in mitigating overfitting on noisy annotated samples.
4.5 Visualization

4.5.1 Confusion Plots

In addition to the overall accuracy, we also analysed the individual expression recognition performance of DNFER on benchmark FER datasets RAFDB, FERplus and AffectNet (shown in Fig. 11) as well as on real-noisy annotated AffectNet-7 (shown in Fig. 12). Clearly, Happiness is the easiest expression to recognize across all the datasets followed by Neutral, Surprise and Anger whereas Fear and Disgust are difficult expressions. However, contempt is the most difficult expression in FERPlus, exhibiting high degree of confusion with neutral. On real noisy annotated AffectNet-7, Neutral, Happiness and Surprise are easy expressions whereas Anger and Disgust are difficult ones. Disgust is often confused with Anger whereas Anger is confused with Neutral. Overall, DNFER performs consistently in discriminating
expressions in synthetic and real noisy annotated datasets as well as on clean datasets. DNFER exhibits superior or comparable performance on all the benchmark FER datasets.

4.5.2 t-SNE plots

In order to evaluate the effectiveness of proposed DNFER method, we use t-SNE [45] to visualize the facial features distribution extracted in 2-D space by Baseline in Fig. 13 (a) and (c) on RAFDB and FERPlus respectively. These are compared with corresponding feature distribution learnt using DNFER in Fig. (b) and (d) respectively. We can observe that facial expression features using Baseline are not discriminating due to large intra-class variation as well as inter-class variation. In contrast, expression features using DNFER have high intra-class similarity and inter-class differences, thus giving superior performance.

4.6 Comparison with the State-of-the-art

We compare proposed DNFER method with existing SOTA methods on 4 popular in-the-wild FER benchmark datasets, RAFDB (see Table 3), FERPlus (see Table 5), SFEW (see Table 4) and AffectNet (see Tables 6 and 7) datasets. Our DNFER obtains top performance on RAFDB (90.41%) compared to other SOTA methods (an improvement by a margin of 0.94 w.r.t top performing FDRL). Although DMUE, RUL, SCN and ULC-AG are designed to deal with uncertain
labels, their performance is inferior compared to DNFER which utilizes all the samples for feature learning. Our DNFER obtains comparable performance on FERPlus (89.32%). On SFEW, DNFER obtains (57.8%) same performance as that of recent top performing model LTLAttenNet. It achieves competitive performance on AffectNet-8 (61.6%) and Affectnet-7 (65.22%) compared to recent SOTA methods. These remarkable results demonstrate the effectiveness of proposed DNFER to deal with noisy annotations present in real-world FER datasets.

Table 3: Comparison on RAFDB (* denotes FER methods with noisy labels robust learning)

| Method               | Year | Accuracy (%) |
|----------------------|------|--------------|
| RAN [4]              | 2020 | 86.90        |
| IPA2LT [34]          | 2020 | 86.77        |
| THIN [46]            | 2022 | 87.81        |
| Mahmoudi et al. [47] | 2020 | 88.02        |
| Liu et. [48]         | 2020 | 88.02        |
| SCN* [15]            | 2020 | 88.14        |
| OADN [1]             | 2020 | 89.83        |
| GCN [49]             | 2020 | 89.41        |
| SCAN [2]             | 2021 | 89.02        |
| FER-VT [50]          | 2021 | 88.26        |
| DMUE* [19]           | 2021 | 88.76        |
| RUL* [31]            | 2021 | 88.98        |
| EfficientFace [6]    | 2021 | 88.36        |
| FDRL [5]             | 2021 | 89.47        |
| ULC-AG* [17]         | 2022 | 89.31        |
| **DNFER**            |      | **90.41**    |

Table 4: Comparison on SFEW

| Method               | Year | Accuracy (%) |
|----------------------|------|--------------|
| IdentityCNN [31]     | 2019 | 50.98        |
| Island loss [52]     | 2018 | 52.52        |
| RAN [4]              | 2020 | 56.40        |
| DMUE* [19]           | 2021 | 57.12        |
| LNLAttenNet [3]      | 2022 | 57.8         |
| **DNFER**            |      | **57.77**    |

5 Conclusion

In this paper, we propose a novel and effective Dynamic Adaptive Threshold (DNFER) framework to learn expression discriminating features robust to noisy annotations. DNFER selects clean samples based on dynamic class-specific threshold for supervised learning, thus mitigating the influence of noisy annotations samples to prevent overfitting as well as imbalance of classes in terms of intra-class difficulty and inter-class sample size. In addition, it uses all the samples by aligning the posterior distributions of weakly-augmented and strongly-augmented image using a consistency loss for expression feature representation. Extensive experiments on synthetic as well as real noisy annotated FER datasets demonstrate its effectiveness in handling noisy annotations. Further, SOTA performance on popular benchmark
Table 5: Comparison on FERPlus (* denotes FER methods with noisy labels robust learning)

| Method                | Year | Accuracy (%) |
|-----------------------|------|--------------|
| Georgescu et al.[53]  | 2019 | 87.76        |
| RAN [4]               | 2020 | 88.55        |
| ESR [54]              | 2020 | 87.15        |
| OADN [1]              | 2020 | 88.71*       |
| SCN* [15]             | 2020 | 88.01        |
| GCN [49]              | 2020 | 89.39        |
| SCAN [2]              | 2021 | 89.42        |
| DMUE* [19]            | 2021 | 88.64        |
| LNLAttenNet [3]       | 2022 | 86.15        |
| **DNFER**             |      | **89.32**    |

Table 6: Comparison on AffectNet-7

| Method                | Year | Accuracy (%) |
|-----------------------|------|--------------|
| Annotators Agreement [10] | 2017 | 65.3         |
| HERO [55]             | 2019 | 62.11        |
| Yongjian et al. [56]  | 2019 | 62.7         |
| FMPN [57]             | 2019 | 61.5         |
| Georgescu et al.[53]  | 2019 | 63.31        |
| IPA2LT [34]           | 2020 | 57.31        |
| THIN [46]             | 2020 | 63.97        |
| OADN [1]              | 2020 | 64.06        |
| GCN [49]              | 2020 | 64.46        |
| SCAN [2]              | 2021 | 65.14        |
| **DNFER**             |      | **65.22**    |

Table 7: Comparison on AffectNet-8 (* denotes FER methods with noisy labels robust learning)

| Method                | Year | Accuracy (%) |
|-----------------------|------|--------------|
| Georgescu et al. [53] | 2019 | 59.58        |
| RAN [4]               | 2020 | 59.50        |
| ESR [54]              | 2020 | 59.3         |
| SCN [15]              | 2020 | 60.23        |
| GCN [49]              | 2020 | 60.58        |
| SCAN [2]              | 2021 | 61.73        |
| ULC-AG [17]           | 2022 | 61.57        |
| DMUE* [19]            | 2021 | 62.84        |
| LNLAttenNet [3]       | 2022 | 59.28        |
| **DNFER**             |      | **61.67**    |

FER datasets validates the utility of DNFER as a general purpose FER method. In future, DNFER can be extended to handle the task of AU prediction and valence-arousal estimation in multi-task setting.

6 Acknowledgments

We dedicate this work to Our Guru and Guide Bhagawan Sri Sathya Sai Baba, Divine Founder Chancellor of Sri Sathya Sai Institute of Higher Learning, Prasanthi Nilayam, Andhra Pradesh, India.

References

[1] Hui Ding, Peng Zhou, and Rama Chellappa. Occlusion-adaptive deep network for robust facial expression recognition. In 2020 IEEE International Joint Conference on Biometrics (IJCB), pages 1–9. IEEE, 2020.

[2] Darshan Gera and S Balasubramanian. Landmark guidance independent spatio channel attention and complementary context information based facial expression recognition. Pattern Recognition Letters, 145:58–66,
A PREPRINT - August 23, 2022

[3] Shasha Mao, Guanghui Shi, Shuiping Gou, Dandan Yan, Licheng Jiao, and Lin Xiong. Adaptively lighting up facial expression crucial regions via local non-local joint network. *arXiv preprint arXiv:2203.14045*, 2022.

[4] Kai Wang, Xiaojiang Peng, Jianfei Yang, Debin Meng, and Yu Qiao. Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Transactions on Image Processing*, 29:4057–4069, January 2020.

[5] Delian Ruan, Yan Yan, Shenqi Lai, Zhenhua Chai, Chunhua Shen, and Hanzi Wang. Feature decomposition and reconstruction learning for effective facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7660–7669, 2021.

[6] Zengqun Zhao, Qingshan Liu, and Feng Zhou. Robust lightweight facial expression recognition network with label distribution training. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 3510–3519, 2021.

[7] Shan Li and Weihong Deng. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Transactions on Image Processing*, 28(1):356–370, 2019.

[8] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2584–2593. IEEE, 2017.

[9] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang. Training deep networks for facial expression recognition with crowdsourced label distribution. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, page 279–283, 2016.

[10] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 2017.

[11] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 2106–2112. IEEE, 2011.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[13] Devansh Arpit, Stanis Jastrz, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International Conference on Machine Learning*, pages 233–242. PMLR, 2017.

[14] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM*, 64(3):107–115, feb 2021.

[15] Kai Wang, Xiaojiang Peng, Jianfei Yang, Shijian Lu, and Yu Qiao. Supressing uncertainties for large-scale facial expression recognition. In *CVPR*, pages 6897–6906, 2020.

[16] Darshan Gera, Vikas G N, and Balasubramanian S. Handling Ambiguous Annotations for Facial Expression Recognition in the Wild. Association for Computing Machinery, New York, NY, USA, 2021.

[17] Yang Liu, Xingming Zhang, Janne Kauttonen, and Guoying Zhao. Uncertain label correction via auxiliary action unit graphs for facial expression recognition. *arXiv preprint arXiv:2204.11053*, 2022.

[18] Shasha Mao, Guanghui Shi, Licheng Jiao, Shuiping Gou, Yangyang Li, Lin Xiong, and Boxin Shi. Label distribution amendment with emotional semantic correlations for facial expression recognition. *arXiv preprint arXiv:2107.11061*, 2021.

[19] Jiahui She, Yibo Hu, Hailin Shi, Jun Wang, Qiu Shen, and Tao Mei. Dive into ambiguity: Latent distribution mining and pairwise uncertainty estimation for facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6248–6257, 2021.

[20] Darshan Gera and S Balasubramanian. Noisy annotations robust consensual collaborative affect expression recognition. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 3578–3585, 2021.

[21] Xingyu Fan, Zhongying Deng, Kai Wang, Xiaojiang Peng, and Yu Qiao. Learning discriminative representation for facial expression recognition from uncertainties. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 903–907. IEEE, 2020.

[22] Xinshao Wang, Yang Hua, Elyor Kodirov, and Neil M Robertson. Imae for noise-robust learning: Mean absolute error does not treat examples equally and gradient magnitude’s variance matters. *arXiv preprint arXiv:1903.12141*, 2019.
[23] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 322–330, 2019.

[24] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1944–1952, 2017.

[25] Jacob Goldberger and Ehud Ben-Reuven. Training deep neural-networks using a noise adaptation layer. ICLR, 2017.

[26] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Lijia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In International Conference on Machine Learning, pages 2304–2313. PMLR, 2018.

[27] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor W Tsang, and Masashi Sugiyama. Co-teaching: robust training of deep neural networks with extremely noisy labels. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, pages 8536–8546, 2018.

[28] Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? In International Conference on Machine Learning, pages 7164–7173. PMLR, 2019.

[29] Hongxin Wei, Lei Feng, Xiangyu Chen, and Bo An. Combating noisy labels by agreement: A joint training method with co-regularization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13726–13735, 2020.

[30] Fahad Sarfraz, Elahe Arani, and Bahram Zonooz. Noisy concurrent training for efficient learning under label noise. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 3159–3168, 2021.

[31] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In International Conference on Machine Learning, pages 4334–4343. PMLR, 2018.

[32] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. IEEE Transactions on Neural Networks and Learning Systems, 2022.

[33] Bodi Yuan, Jianyu Chen, Weidong Zhang, Hung-Shuo Tai, and Sara McMains. Iterative cross learning on noisy labels. In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pages 757–765. IEEE, 2018.

[34] Darshan Gera, S Balasubramanian, and Anwesh Jami. Cern: Compact facial expression recognition net. Pattern Recognition Letters, 155:9–18, 2022.

[35] Jiabei Zeng, Shiguang Shan, and Xilin Chen. Facial expression recognition with inconsistently annotated datasets. In Proceedings of the European conference on computer vision (ECCV), pages 222–237, 2018.

[36] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In International Conference on Machine Learning, pages 4334–4343. PMLR, 2018.

[37] Eran Malach and Shai Shalev-Shwartz. Decoupling" when to update" from" how to update". Advances in Neural Information Processing Systems, pages 960–970, 2017.

[38] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In Proceedings of the eleventh annual conference on Computational learning theory, pages 92–100, 1998.

[39] Darshan Gera and S Balasubramanian. Co-curing noisy annotations for facial expression recognition. ICTACT Journal on Image and Video Processing (IJIVP), 2021.

[40] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pages 702–703, 2020.

[41] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hammer, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in representation learning: A report on three machine learning contests. In International conference on neural information processing, pages 117–124. Springer, 2013.

[42] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters, 23(10):1499–1503, 2016.
[44] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. *ECCV*, 2016.

[45] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[46] Estephe Arnaud, Arnaud Dapogny, and Kevin Bailly. Thin: Throwable information networks and application for facial expression recognition in the wild. *IEEE Transactions on Affective Computing*, 2022.

[47] M Amine Mahmoudi, Aladine Chetouani, Fatma Boufera, and Hedi Tabia. Kernelized dense layers for facial expression recognition. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 2226–2230. IEEE, 2020.

[48] Ping Liu, Yuewei Lin, Zibo Meng, Lu Lu, Weihong Deng, Joey Tianyi Zhou, and Yi Yang. Point adversarial self-mining: A simple method for facial expression recognition. *IEEE Transactions on Cybernetics*, 2021.

[49] Ping Jiang, Bo Wan, Quan Wang, and Jiang Wu. Fast and efficient facial expression recognition using a gabor convolutional network. *IEEE Signal Processing Letters*, 27:1954–1958, 2020.

[50] Qionghao Huang, Changqin Huang, Xizhe Wang, and Fan Jiang. Facial expression recognition with grid-wise attention and visual transformer. *Information Sciences*, 580:35–54, 2021.

[51] Debin Meng, Xiaojiang Peng, Kai Wang, and Yu Qiao. Frame attention networks for facial expression recognition in videos. In *2019 IEEE international conference on image processing (ICIP)*, pages 3866–3870. IEEE, 2019.

[52] Jie Cai, Zibo Meng, Ahmed Shehab Khan, Zhiyuan Li, James O’Reilly, and Yan Tong. Island loss for learning discriminative features in facial expression recognition. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 302–309. IEEE, 2018.

[53] Mariana Iuliana Georgescu, Radu Tudor Ionescu, and Marius Popescu. Local learning with deep and handcrafted features for facial expression recognition. *IEEE Access*, 7:64827–64836, 2019.

[54] Henrique Siqueira, Sven Magg, and Stefan Wermter. Efficient facial feature learning with wide ensemble-based convolutional neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 5800–5809, 2020.

[55] Wentao Hua, Fei Dai, Liya Huang, Jian Xiong, and Guan Gui. Hero: Human emotions recognition for realizing intelligent internet of things. *IEEE Access*, 7:24321–24332, 2019.

[56] Yongjian Fu, Xintian Wu, Xi Li, Zhijie Pan, and Daxin Luo. Semantic neighborhood-aware deep facial expression recognition. *IEEE Transactions on Image Processing*, 29:6535–6548, 2020.

[57] Yuedong Chen, Jianfeng Wang, Shikai Chen, Zhongchao Shi, and Jianfei Cai. Facial motion prior networks for facial expression recognition. *VCIP*, 2019.