Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2023.0322000

DEER: Deep Emotion-sets for fine-grained Emotion Recognition

Sharjeel Tahir[1], Nima Mirnateghi[1], Syed Afaq Ali Shah[1], and Ferdous Sohel[2]
[1]School of Science, Edith Cowan University, Joondalup, WA 6027 Australia
[2]School of Information Technology, Murdoch University, Murdoch, WA 6150 Australia

Corresponding author: Sharjeel Tahir (e-mail: s.tahir@ecu.edu.au).

ABSTRACT For robots to effectively interact with humans in-the-wild, it is essential that they accurately recognize their emotions. To achieve this, important facial features must be captured to reliably comprehend human emotions. Most facial emotion recognition (FER) research works have used single-shot images for classifying emotions, and in certain instances, several networks have been utilized for voting against each image. These approaches have functioned well; however, there is potential for improvement in terms of precision. In this paper, we propose emotion-sets as a unique encoding for face image data (with various people and face angles) to classify emotion classes, as opposed to the conventional single-image-based classification. For each image in an emotion-set, prediction confidence against each emotion is utilized as a vote. The results are generated by a combination of two distinct voting methods, including Majority Voting and Weighted Voting. Without data augmentation, feature extraction, or additional training data—techniques that the majority of state-of-the-art works have used—the proposed technique achieves state-of-the-art accuracy on the Facial Emotion Recognition 2013 (FER2013), Cohn Kanade (CK+), and Facial Emotion Recognition Group (FERG) datasets. Our experimental findings indicate that the suggested emotion-set classification yields more accurate results than the current state-of-the-art FER methods.

INDEX TERMS Emotion Recognition, Facial Emotion Recognition, Image set classification, Majority Voting, SFEW, VGGFace, Weighted Voting.

I. INTRODUCTION

In the growing world of human-computer interaction and artificial intelligence, there has been an ever-increasing boom in the number of methodologies for making the machines more and more intelligent, human-friendly, and flexible to different environments. This being said, there are still many challenges hindering the way these machines interpret human emotions. The challenges lie, broadly, in the way machines perform the analysis of speech, facial expressions, and body gestures [1]. The dominant challenges in the task of emotion recognition include, small amount of publicly available data, variation in age of the participants, pose and illumination [2]. In order to understand human emotions, it is of key importance to identify the emotions of people e.g., from their facial expressions, voice tone, and gestures.

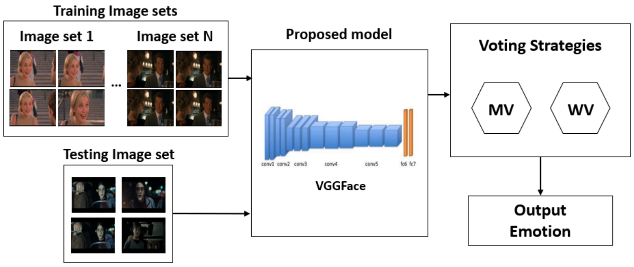

FIGURE 1. **Pipeline of the proposed facial emotion recognition technique**

Literature review of image based emotion recognition suggests that many recent studies have focused on deep learning based methods. Although a bulk of the research has been taken out using Convolutional Neural Network (CNN) models [3], some researchers have incorporated hybrid CNN-RNN (Recurrent Neural Networks) methods [4]. The topic has gained more interest since the launch of the EmotiW challenge in 2013. In the subsequent years, the challenge came with a new problem every year. Besides, researchers have also surpassed the previous best results of these challenges [5], with the help of adequate data, and new powerful network architectures and algorithms. For instance, in the last EmotiW Challenge, held in 2016, Fan et al. [4] improved the precision of previous winner's results by approximately 6 percent, by combining 3D Convolutional Networks (C3D) with RNN in a late-fusion method. The most effective model used for the Active Facial Emotion in the Wild (AFEW) video-based dataset is the TMSAU-Net [6]. However, the challenges of pose variance and illumination, age, and camera angle dissimilarity still limit the accuracy to a maximum of 47.6 percent, to date.

Another strategy that has proven to be more effective than others is the shifting from single-shot-based image classification to image set classification. Image set classification involves the formation of sets within the given dataset, followed by making predictions on the basis of similarity between the image sets. There are various reasons why this technique has proven to be better. Firstly, an image set offers multiple variations in data with respect to, e.g., pose, illumination, occlusion, and view point for a given subject [7]. Secondly, because of the number of samples per subject, it helps manifest diverse hidden features for the algorithms to learn [8].

Presented in the thesis [9], in this paper, we propose a deep learning based image set classification technique for facial emotion recognition. We use two different voting strategies on the testing image sets i.e., majority voting and weighted voting. Evaluation of the proposed technique is performed on four publicly available datasets: FER2013 [10], FERG [11], CK+ and SFEW [12].

The major contributions of this work are as follows:

- A novel deep learning based image set classification for the task of FER has been proposed. To the best of our knowledge, this is the first ever technique.
- Our proposed technique outperforms the state-of-the-art methods on FER2013, CK+ and FERG datasets with the highest accuracy of 100%. In addition, we achieve superior results on the SFEW dataset as well.

The rest of the paper is organised as follows: In Section II, we discuss related work. Section III describes our proposed methodology. Section IV reports the experimental results and analysis. The conclusion is provided in Section V.

II. LITERATURE REVIEW

A setback in the older works is the usage of staged data i.e., data was collected in controlled environment. To resolve this, [12] introduced a challenging and spontaneous dataset extracted from real-world clips. This dataset emerged from the Emotion Recognition in the Wild (EmotiW) challenge. The challenge brought advancements in the task of FER because of its challenging conditions – as the name "in the wild" refers to "uncontrolled and real-life data". The datasets predominantly used for this challenge over the years are AFEW/SFEW. These datasets were gathered from movies that pose scenes close to the real-world conditions [13].

In addition to the availability of advanced datasets, the focus has now shifted from only audio or image based unimodal approach to multi-modal approaches that use both visual and speech data [14], or techniques that involve concatenation of diversified features [15] [16] for better results. For further information, a briefly drafted comparison of these approaches can be found in these review papers on latest trends in ER [17] and [18].

With the variations in modalities, researchers have also tried to increase the spectrum of understanding of emotions by going beyond the seven basic emotions. These basic emotions are Angry, Happy, Sad, Neutral, Disgusted, Shocked and Fearful. This was pioneered by Du et al. [19] in 2014 introducing CEs (Compound Emotions). Their proposed method does classification amongst 22 distinct emotions formed on the basis of basic emotions, such as sadly angry, happily surprised and vice versa. The work deduces that most of the hybrid categories of emotions are visually distinguishable and they can, hence, give a better understanding of human emotions. This work has by far been one of the best contributions in extended emotion categories.

Another variation in the field can be seen where a versatile approach is used by Vithanawasam et al. [20], recognizing emotions by using both facial expressions and upper-body gestures. The work has a good performance with an average of 75% accuracy for three basic emotions i.e. anger, fear, and bored. L.B et al. [21] used a hybrid methodology for recognition of emotions through eye and head movement of subjects and estimating their concentration levels accordingly.

Emotion recognition using deep learning methods have a long history. In order to get a better understanding of the topic, we will need to first discuss conventional techniques, which have been deployed for emotion recognition related applications. For a detailed review of emotion recognition approaches, readers are referred to a survey by Imani et al. [22], as it presents a detailed discourse analysis along with statistical comparison of most of the popular ER methods.

A. CONVENTIONAL TECHNIQUES FOR ER

The basic pipeline of facial emotion recognition techniques broadly consists of three stages; facial image acquisition, feature extraction, and classification. The methods mainly used were landmark predictors, and facial detection by top-down or bottom-up knowledge-based techniques [23]. For feature extraction, HOG (Histogram of Gradients), LBP (Local Binary Patterns) and Euclidean Distance were largely used. Classification tasks were mainly carried out by KNN, Bayesian classifiers, or SVM (Support Vector Machines). These methods, however, require manual selection of features, which was eliminated later on by DL methods.

Some of the works that use the conventional techniques for FER are discussed hereinafter. Shojaeilangari et al. [24] proposed a pose-invariant OF (Optical Flow)-based spatio-temporal descriptor for representation of facial emotions that is robust to extreme pose variations, i.e., head movements. They put forward a dictionary-based classification method, i.e., Extreme Sparse Learning (ESL). Also, a supervised sparse coding algorithm called Class Specific Matching Pursuit (CSMP) was proposed, the motivation behind which was taken from the simultaneous sparse approximation algorithm and Simultaneous Orthogonal Matching Pursuit (S-OMP). The work by Kirana et al. [25] proposes emotion recognition based on the famous Viola Jones algorithm. Although the Viola Jones algorithm is used specifically for face detection, yet their work exploits it for both detection of face and emotion recognition, using the rectangular feature and cascading Adaboost algorithm.

Experimentation is performed on two bimodal datasets, i.e., eNTERFACE '05 and AFEW 4.0. In a work by Ding et. al., [26] peak expression frames were extracted using Double Local Binary Pattern (DLBP). In the next step, facial expression recognition is performed on two datasets using Taylor expansion theorem. They also empirically show how their proposed model's performance is comparable to a neural network, i.e., AlexNet. However, the time complexity of their model is still far higher than AlexNet. An implementation of SVMs for emotion recognition was proposed by Hong et al. [27], where the model consists of a novel Support Vector Pursuit Learning (SVPL) to minimize training over old facial data and to improve the process of retraining the SVMs.

B. DEEP LEARNING TECHNIQUES FOR ER

A detailed report on the chronological progression and evolution of the deep learning methods can be found in the survey paper [28]. Initially, attempts were made to recognize emotions only using CNNs [29]. Later, Concatenation of multiple networks was performed by Jung et al. [30], who integrated a fully-connected DNN – Deep Temporal Geometry Network (DTGN) with a CNN in order to extract the temporal features from data.

It is well established from a variety of studies [31] [22] that a successful way to perform emotion recognition is by forming an ensemble of deep learning networks, especially CNNs and RNNs. Perrone et al. [32], explain how an ensemble of algorithms can minimize square errors and can also avoid over-fitting in the training, along with being capable of parallel computation for training and testing. Implementing ensemble of networks, Wang et al. [33] proposed an Oriented Attention Enable Network (OAENet) network for FER.

Zhou et al. [34] proposed the Feature Refinement (FeatRef) network to learn expression specific features. They present state-of-the-art performance on existing Micro Expression Recognition (MER) datasets.

Many researchers have proposed to use facial Action Units (AU) i.e. AU occurrence and AU intensity for FER. Liu et al. [35] used a three-layered model where the first layer generates a complete representation using the convolution and max-pooling layers. In the subsequent layer, Action Unit Receptive Fields (AURFs) are brought about using feature selection scheme for identifying local appearance variations. In the last layer, multi-layer Restricted Boltzmann Machines (RBM) are used for hierarchical features leading to FER. Introducing a new facial expression dataset FG-emotion and giving state-of-the-art result performance on SFEW, Liang et al. [6] proposed an end-to-end Multi-Scale Action Unit (AU)-based Network (MSAU-Net), along with detailed performance benchmark on several image and video based datasets. In another work, a deep learning-based approach was put forth by Liliana et al. [36]. They introduced an automatic feature extraction system using deep learning convolutional neural network, and secondly, employing a comprehensive dataset; CK+ Database. Exploring the adversarial domain, Bozorgtabar et al. [37] exploited the use of GANs to regenerate better quality images and use them to perform facial expression recognition on new data without the need to retrain the model on the target images.

Sheng et al. [38] proposed a gait-based joint identity and expression recognition system using the pose data. They also presented a new dataset - EMOGAIT containing 1440 labelled video sequences. However, emotions were limited to four categories i.e., happy, angry, sad and neutral. In 2019, Erol et al. [39] performed multimodal emotion recognition in a Human Robot Interaction (HRI) scenario experimentation using CNN with LSTMs. Wang et al. [40] presented a domain adaptation SPD matrix network to recognize emotional representation of people across three categories i.e., Valence, Arousal and Dominance using EEG signals.

From the aforementioned studies, it is evident that the problem of lower accuracy faced with the conventional methods is ameliorated by deep learning models [22].

C. IMAGE SET CLASSIFICATION

In image set classification, multiple images of a given subject are grouped together to form sets. These sets are then used for the training and testing of the model [8]. It enables the model to capture additional information while providing robustness against issues such as, occlusion, pose variance, illumination and other variations within the images. Because there are deeper features to be extracted from multiple images of each subject i.e., variations in poses and illumination, the model can generate higher correlation between images of similar classes.

There are mainly three types of approaches to perform image set classification, namely: parametric, non-parametric and deep learning methods.

1) Parametric Methods

Parametric methods use statistical models to represent image sets after which the estimation is done on the basis of difference between the probabilities of image sets on a plot. Only a few works have been done using parametric methods because of their inability to perform well when the training and testing sets are not much similar to each other. The work by Lee et al. [41] uses a manifold to represent and determine pose of the objects. Similarly, Kim et al. [42] proposes a semi-parametric method, where the probability density functions are modelled as Gaussian Mixture Models (GMMs) on low-dimensional non-linear manifolds, and the similarity between the predicted densities is evaluated using Kullback-Leibler divergence. The research work by Yamaguchi et al. [43] represents each subject (person) by a manifold, which is approximated on an affine plane. Multiple clusters are formed using K-means algorithm, where each cluster is represented as a plane.

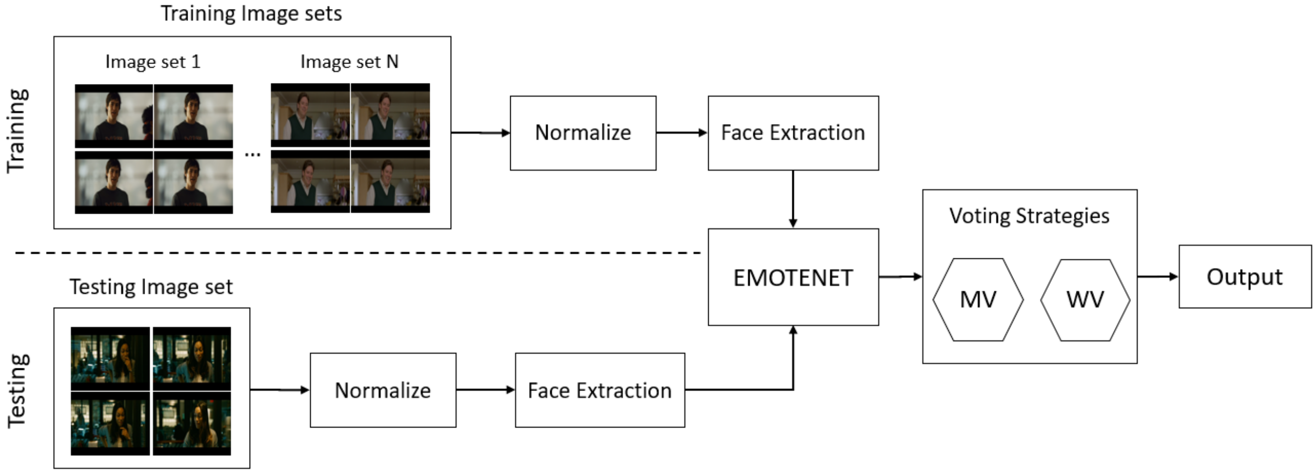

FIGURE 2. Proposed Framework. Training and Testing image sets are both pre-processed using the same techniques before passing to the deep neural network (EMOTENET). Voting techniques i.e., Majority Voting and Weighted Voting are used to classify each image from the testing image sets.

2) Non-Parametric Methods

Non-parametric methods are different from the parametric in such a way that they do not rely on statistical modelling. To mention a few major works; [44] is inspired by an LDA (Linear Discriminant Analysis) system, and the classes are predicted by comparing canonical correlations within image sets. Following this, AHISD (Affine Hull based Image Set Distance) is used to compare the similarity between image sets. Similarly, [45] represents images as points and image sets as convex geometric region, where predictions are made on the basis of geometric distances between the convex hulls. Accelerated Proximal Gradient Projection (APGP) algorithm is proposed by Mian et al. [46] to project the similarity between image sets being represented as convex regions.

These methods, however, perform on-to-one image set comparison one-to-one making it computationally expensive. Therefore, it is not feasible when the number of image sets is high.

3) Deep Learning Methods

There is a relatively small body of literature that performs image set classification using deep learning techniques. To begin, an extensive work by Hayat et al. [8] proposed a Template Deep Reconstruction Model (TDRM) based framework, where unsupervised learning of the model is performed using Gaussian Restricted Boltzmann Machines (GRBMs). The basic structure of their model is based on an auto-encoder that regenerates input images. For evaluation, three different voting strategies have been used i.e. MV, WV, Preferential Weighted Voting (PWV).

Shah et al. [47] proposes an Iterative Deep Learning Model (IDLM), where discriminative representations from raw facial images can be automatically and hierarchically learned using the model. To learn the low-level features from the data, a Pooled Convolutional Layer (PCL) has been proposed. On the other hand, Artificial Neural Networks (ANNs) have been implemented in a hierarchical way, for an unsupervised feature learning of the input image sets.

In another work, Shah et al. [48] proposed image set classification utilizing linear regression models. Sub-spaces of high-dimensional space are projected as image sets in a gallery. Whereas, each test image is reconstructed with the help of regression models. Classification is performed on the basis of minimum calculated error between reconstructed and original images. Similar to the previously discussed study [8], this work also uses voting strategy for evaluation, however, only one strategy i.e WV, was used in this method.

A contemporary work in this area was put forward by Wang et al. [49]. They propose a Symmentric Positive Definite (SPD) manifold deep learning network (SymNet) for image set classification. 2-D Principal Component Analysis technique is adapted in order to minimize the complexity of computational functions for building and training, while allowing the network to learn the multistage weights, efficiently. The network is followed by the Kernel Discriminant Analysis (KDA) in combination with the output vectorized feature representation for subspace learning.

It is evident from the previously mentioned works that image set classification is more feasible. However, there are only a handful amount of works that have incorporated deep learning techniques with image set classification. Moreover, to the best of our knowledge, there is no existing study, that uses deep learning methods with image set classification for the task of facial emotion recognition.

III. PROPOSED METHODOLOGY

A block diagram of the proposed technique is shown in Figure 2.

The proposed deep learning framework can be divided into four stages i.e., input, face extraction, training and testing the network on image sets using different voting strategies. We use the Dlib face extraction method for the extraction of facial features. It is carried out only on the SFEW dataset [12]. As for the training, we implement the VGGFace Network proposed by [50], with a few changes in the architecture (details in Section III-H), as per the requirement of the datasets under examination. After this, the trained network is used to determine emotions using the image sets. The image sets are formed mainly per person in each gallery. The next phase is where the predicted classes are put together for voting. We use two different strategies for voting: MV and WV. In the following, we discuss the different modules of our proposed framework.

A. FACE EXTRACTION

In order to perform emotion recognition from faces, it is necessary to extract facial features, which is a challenging task, especially when the data contains noise, occlusion, pose variance and other challenges. To overcome such challenges, we used Dlib face extraction toolkit[1]. This library is specifically designed for object detection/extraction, and has a better precision in comparison to other methods when training on challenging data. We deploy the face extraction part particularly on the SFEW dataset, as this dataset comprises of images with occlusion, pose and illuminance variations, and cultural/age differences [5].

B. PRE-PROCESSING

The first step performed in the data pre-processing is to normalize the images. Let I be the input image. The normalized image can be given as:

$$I_N = \{(I - \alpha)/(\beta - \alpha)\} \tag{1}$$

where α is the minimum pixel value, and β represents the highest pixel value.

C. OVER SAMPLING (DATA AUGMENTATION)

As discussed previously in Section III-A, the SFEW and CK+ datasets pose the challenge of poor balance between the classes and a very low number of images. To this end, we deploy the technique of over sampling, where we perform data augmentation on the two datasets (SFEW and CK+) that have a lower number of samples, to generate synthetic images.

The data augmentation is performed disparately for each dataset, according to degree of class-imbalance. We select four augmentation techniques: rotation of 20 degrees, random zoom of 10 percent, horizontal flip, and vertical flip.

D. IMAGE SET FORMULATION

Let X is an image set that contains multiple images M, where the number of images within an image set is T.

$$X_i = \{M_1, M_2, M_3...., M_T\} \tag{2}$$

where $i = \{1, 2, 3, ..., N\}$. Similarly, a gallery G, where the total number of image sets is N can be represented as;

$$G_\triangle = \{X_1, X_2, X_3, ..., X_N\} \tag{3}$$

where $\triangle = \{1, 2, 3, ..., \delta\}$. Therefore, the total number of images in a gallery can be given as;

$$G = \{M_1, M_2, M_3, ..., M_{N*T}\} \tag{4}$$

E. TRAINING

To train the dataset, we combine all images in each gallery to form one set, where a gallery can have more than one image sets. Our training algorithm is shown in Algorithm 1.

$$D_{train} = \{G_1, G_2, G_3, ..., G_\delta\} \tag{5}$$

where δ is the total number of classes, and D_{train} is the training data.

Training using Image Sets

$G_\triangle$ in D_{train} $X_M \notin G_\triangle$ Create N number of image sets X_N for $G_\triangle$ **Input:** Training image sets X_{train} Normalise $\forall M \in X_{train}$ Extract Faces $\forall M \in X_{train}$ Augment $\forall M_{ext} \in X_{train}$ Train $EMOTENet(X_N)$ **Output:** $EMOTENet$

F. TESTING

Given X_{test} is an image set containing images $M_1, M_2, M_3, ..., M_T$, that belongs to the testing gallery G_{test}. Therefore the model prediction for the testing image set X_{test} can be calculated as:

$$P_{X_{test}} = \{P_{M_1}, P_{M_2}, P_{M_3}, ..., P_{M_T}\} \tag{6}$$

where $P_{X_{test}}$ is a set of probability based predictions made against each image M_j, where $j = \{1, 2, 3, ..., T\}$ in the testing image set X_{test}.

Emotion recognition using image set classification

Input: X_{test}^N each gallery $G_{test(\triangle)}$ in D_{test} each image set X_{test} in $G_{test\triangle}$ Normalise X_{test} $P_{X_{test}} =$ Predict(X_{test}, $EMOTENet$) See Eq. (6) MV: $\hat{MV}_{X_{test}} = mode(Y_{test})$ See Eq. (7-8) WV: $\hat{Y}_{X_{test}} = \arg\max(WV_{X_{test}})$ See Eq. (9-11) **return** Label y_{test} of X_{test}

G. VOTING STRATEGIES

In order to evaluate the overall predicted probability $P_{X_{test}}$ against each test image set, we devise two voting strategies i.e. MV and WV.

1) Majority Voting

In majority voting, each image M casts a vote V_M on the basis of maximum probability P_M for the given image,

$$V_M = argmaxP_M(M) \tag{7}$$

In the subsequent step, the votes casted by each image M, of the testing image set X_{test}, are compared. The class with the highest number of votes is declared as the nominated class y_{test} for the image set.

[1] www.dlib.net

$$MV_{X_{test}} = mode(Y_{test}) \tag{8}$$

Where Y_{test} is the predicted class for each test image set X_{test}.

2) Weighted Voting

In Weighted Voting, each image M casts a vote for all classes. The vote is then assigned with a weight as per its probability P_M. Let V_w is the vote for an image, where β is a constant, it can be given as:

$$V_{w(M_T)} = e^{-\beta P_{M_T}} \tag{9}$$

Hence, the predicted weighted vote WV for the testing image set X_{test} can be deduced as;

$$WV_{X_{test}} = \sum_{M=1}^{T} V_{w(M_T)} \tag{10}$$

The predicted class for a given test image set is the class with the highest weighted vote, and it can be given as;

$$y_{X_{test}} = argmax(WV_{X_{test}}) \tag{11}$$

H. EMOTENET: EMOTIONALLY INTELLIGENT NETWORK

We adapt the VGG16 [51] network as the backbone of our network architecture. To train our network we use the pre-trained VGGFace weights [52]. VGGFace is a 16-layer CNN that is pre-trained on more than million images of the MS_Celeb_1M dataset. Because the weights are trained on facial features, the performance is higher as compared to other networks. For this reason, the network has shown to be more reliable for the datasets under examination in our work. A detailed analysis of the impact of pre-trained weights on the model's performance is given in Section IV.

To fine-tune the network, the following configurational settings were incorporated. We add two fully connected and dropout layers at the end of our network in order to minimize over-fitting and hence, improve the overall accuracy. The configuration of the added layers is as follows:

First, a 4096 dimensional fully-connected dense layer is integrated next to the backbone network's last pooling layer, followed by a dropout layer with a ratio of 0.5. Then, another fully-connected layer, with the same number of input neurons is added. Activation function for all the added layers is ReLU, so that the computational complexity could be curbed to its minimum.

In addition, L2 regularizer is also used to tackle class imbalance. The network is trained using the Adam optimizer, with a learning rate of 0.0001. A batch size of 16 is used for SFEW, CK+ and FER2013, and 32 for FERG. The number of epochs is set between 8 to 12 for each dataset. Sparse Categorical Crossentropy Loss function is used for training on all the datatsets.

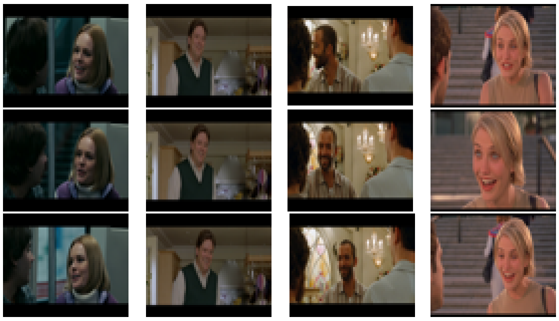

FIGURE 3. Example images from SFEW 2.0 dataset

IV. EXPERIMENTAL ANALYSIS

In this section, we first discuss the datasets used in our experiments and how they are prepared for evaluation. Then, we describe the results of the proposed technique on each dataset in comparison to the state-of-the-art methods. Finally, a detailed ablative study and results are reported for the FER2013 dataset, where we change the training set size and input image resolution, and add noise in the testing data.

A. DATASETS

The proposed technique is implemented across four frequently used publicly available datasets for the task of FER. Emotions are categorised into seven disparate classes including, Anger, Disgust, Fear, Happy, Neutral, Sad, and Surprised.

1) SFEW2.0

Static Facial Emotions in the Wild (SFEW) [12] is an image-based dataset derived from the AFEW dataset that consists of video clips taken from various Hollywood movies. The dataset contains images that are challenging, as they have variations in pose, illumination, occlusion, and focus, as well as a wide range of age and cultural difference. We use the SFEW version 2.0 dataset. The number of images in the training, validation and test sets of the dataset are 958, 436, and 372 respectively. Because the labelled test data is not provided publicly, we perform evaluation on the validation data set, same as the state-of-the-art work [53]. Because of the smaller number of images in each gallery, we perform image augmentation to extend the dataset. Then, the image sets are created as per two strategies i) the given number of images in each gallery, where the number of images per subject are too less and ii) per each person, where the maximum number of images per image set is not above 100 and less than 20. Example images from the dataset are displayed in Figure 3.

2) FER2013

FER2013 is a dataset compiled by Goodfellow [10] with more than 30000 labelled images captured in relatively controlled conditions. Resolution for images is 48*48 and the average number of images per gallery is around 4000. We do not use any augmentation or face extraction for this dataset. Image

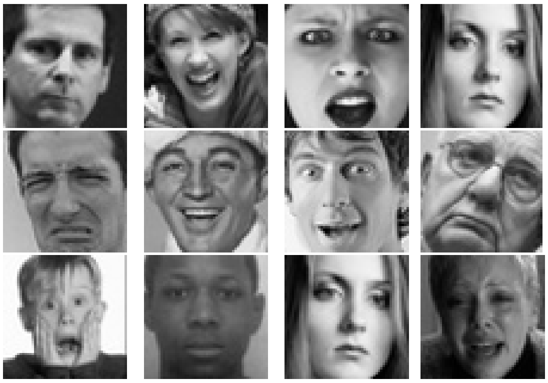

FIGURE 4. Example images from FER2013 dataset

TABLE 1. Accuracy of EMOTENET on all datasets

Dataset	MV	WV
SFEW	52%	56.2%
FER-2013	99.6%	98.1%
FERG	100%	100%
CK+	100%	100%

sets are formed according to the size of gallery, with a maximum of 100 but no less than 30 images per image set. The reason behind this is the variable size of each gallery in the dataset. To exemplify, the number of images in the disgust class is only about 600, whereas, other classes have more than 3000 images each. Example images from the dataset are displayed in Figure 4

3) FERG

The FERG (Facial Expression Research Group-based database) is a dataset comprising of six 3D cartoon characters. For each character, seven emotion galleries are provided. The total number of images provided is 55769. We divide the dataset into two sets i.e., training and test, with approximately 40000 and 15000 images, respectively. In order to form image sets for each emotion class, we first segregate images of each character into emotion based galleries and then divide them into image sets, with each image set containing 180 images. Example images from the datset are dsiplayed in Figure 5.

TABLE 2. Comparison with state-of-the-art methods for SFEW

Method	Pre-trained Dataset	Year	Accuracy
Island Loss	FER2013	2018	52.52%
Identity-aware CNN	FER2013	2017	50.98%
Multiple deep CNNs	FER2013	2015	55.96%
RAN-ResNet18	MS_Celeb_1M	2019	54.19%
RAN(VGG16+ResNet18)	MS_Celeb_1M	2019	56.4%
MSAU-Net	MS_Celeb_1M	2020	**57.4%**
EMOTENET (Ours)	MS_Celeb_1M	2021	56.2%

TABLE 3. Comparison with state-of-the-art methods for FER2013

Method	Year	Accuracy
CNN	2016	62.44%
GoogleNet	2018	65.20%
VGG+SVM	2019	66.31%
Conv+Inception layer	2016	66.40%
Attentional ConvNet	2019	70.02%
CNN+SVM	2013	71.20%
ARM(ResNet18)	2021	71.38%
Inception	2016	71.60%
ResNet	2016	72.40%
VGG+SVM	2016	72.70%
VGGNet	2021	73.28%
Ensemble of 7 CNNs	2019	76.2%
MSAU-Net	2020	78.3%
EMOTENET (Ours)	**2021**	**99.6%**

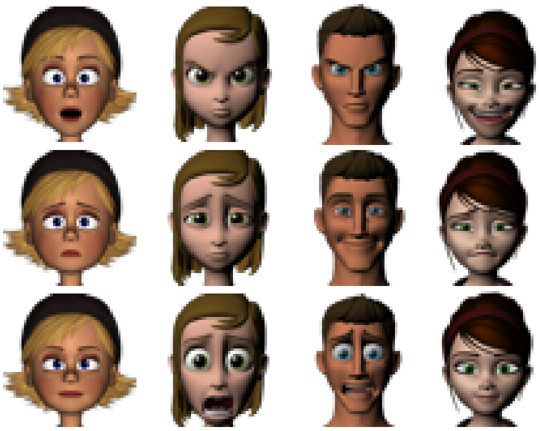

FIGURE 5. Example images from FERG dataset

4) CK+

CK+ consists of 529 videos from 123 subjects, 327 of them annotated with eight expression labels. Each video starts with a neutral expression, and reaches the peak in the last frame. The total number of the images is 1031, which is split into image sets. The image sets for each gallery are sized as per the number of images in the gallery. Size of image sets remain same for the testing and training sets of each gallery. We divide the training and testing data in approximately a 60 to

TABLE 4. Comparison with state-of-the-art methods for FERG

Method	Year	Accuracy
DeepExpr	2016	89.02%
Ensemble Multi-feature	2018	97%
Adversarial NN	2018	98.20%
Attentional CNN	2019	99.30%
EMOTENET (Ours)	**2021**	**100%**

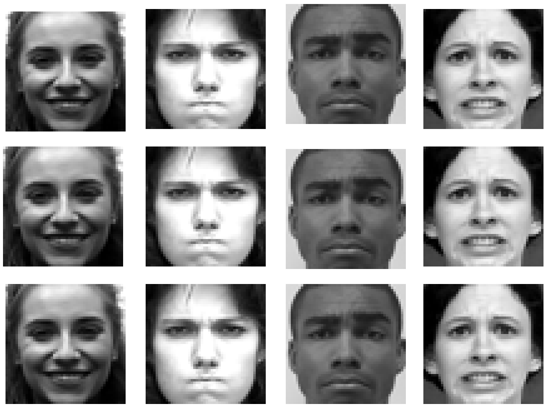

FIGURE 6. Example images from CK+ dataset

40 ratio, where the training and testing sets have 598 and 346 images, respectively. Sample images from the CK+ dataset are depicted in the Figure 6.

TABLE 5. Comparison with state-of-the-art methods for CK+

Method	Year	Accuracy
Liu et al. [54]	2017	97.1%
Ding et al. [55]	2017	98.6%
Yang et al. [56]	2017	97.3%
MSAU Net [6]	2020	99.1%
EMOTENET (Ours)	**2021**	**100%**

B. ANALYSIS

Table 1 reports the accuracy achieved by our technique on all the four datasets. Our proposed technique achieves the state-of-the-art accuracy on the FER2013, CK+ and FERG datasets marking at a highest accuracy of 100 percent for all three datasets. For the SFEW dataset, our architecture achieves 56.2 % accuracy. As reported in Table 1, the weighted voting technique performs better than majority voting, in terms of accuracy, for the SFEW dataset. However, the difference between the accuracy of the two strategies is recorded to be less when the number of images is higher in the test image set. To instantiate, it is evident from Table 1 that the accuracy for FER-2013 and FERG datasets - where the testing image sets have more images, is same for both voting strategies, as opposed to SFEW.

C. COMPARISON WITH STATE-OF-THE-ART METHODS

The current best performance on SFEW2.0 dataset was acheived by Liang et al. [6] in 2020. They achieve an accuracy of 57.2% by deploying an end-to-end model. Yu et al. [5], used an ensemble of multiple deep convolutional neural networks giving an accuracy of 54.19%. The work by [57] uses an Island Loss CNN (IL-CNN) reporting an accuracy of 52.52%. Table 2 gives a summary of the performance and comparison with state-of-the-art works on SFEW.

As for the FER2013 dataset, to the best of our knowledge, the top accuracy benchmarks at 78.3%, achieved by [6]. Using a VGGNet, [58] reported an accuracy of 73.28%. The proposed technique exceeds the state-of-the-art accuracy by a margin, marking at a highest accuracy of 100%. We performed our experiments in 5 folds with different combinations of training and test image sets, reporting an average accuracy of 99.6% for WV and 98% for MV. Importantly, none of the feature extraction or data augmentation techniques was applied. The numerical results are reported in Table 3.

CK+ is similar to the FER2013 dataset with respect to pose and noise in the images. Also, the images contain a limited and same number of people across all emotion classes. This makes it easy to perform FER on this dataset. Hence, the state-of-the-art performance on this dataset is 99.1% by Liang et al. [6]. We outperform the previous works by getting a 100% accuracy using both voting strategies. Comparison of our results for the CK+ datset with the state-of-the-art works is given in Table 5.

The main reason for choosing FERG is that it validates our technique on 3D characters. So far, the best work on this dataset is by Minaee et al. [59] achieving 99.3% accuracy. We have achieved an accuracy of 100% on this dataset with both voting strategies, without using any feature extraction or data augmentation techniques. The results are reported in Table 4.

D. ABLATION STUDY

To validate the generalization ability and to determine the robustness of the proposed method to various challenges, three additional experiments are carried out on FER2013 dataset and reported in this section.

1) Image Noise

To demonstrate the robustness of our proposed technique, we introduce two types of noise i.e., Gaussian noise and Salt & Pepper noise in the test image sets.

For the salt & pepper noise, multiple experiments were performed by alternating the proportion of black and white pixels, and unchanged pixels. The default value is 0 (no black and white pixels). We attempt testing against three disparate values of amount of noise, and the results are shown in Table 6. The accuracy decreases with the increase in the amount of noise percentage. We increase the noise by five percent after each experiment, where the first experiment has 5% noise and the final experiment included 20% noise.

For Gaussian noise, the mean of random distribution is tweaked to five different values as shown in Table 7. When the mean is increased, images get brighter (white noise), however, the decrease in accuracy is not as significant as compared to when the mean is decreased below 0, where the images get darker.

TABLE 6. Accuracy of the proposed technique under Salt & Pepper noise

S. No.	Amount	MV	WV
1	0.05	95.91%	97.27%
2	0.1	91.15%	89.11%
3	0.15	81.63%	78.91%
4	0.2	69.40%	69.40%

TABLE 7. Accuracy of the proposed technique under Gaussian noise

S. No.	Mean	MV	WV
1	0.5	98.63%	95.91%
2	0.1	98.63%	99.31%
3	0	98.63%	99.31%
4	-0.1	99.31%	99.31%
5	-0.25	95.23%	95.23%
6	-0.5	51.20%	48.97%

2) Effect of Image size (Resolution)

The original size of images for this dataset is 48x48. To test our proposed technique against different image resolutions, we increase the size to 56x56, 64x64 and 72x72 pixels, respectively. The resolution is not decreased below 48x48 because our network does not support images below that size. Results are depicted in Table 8. As the original image size is 48x48, up-scaling the image size reduces the quality of images (distorts the details), which in return, limits the performance of the network. Hence, with the increase in image size, a decrease in the accuracy is recorded. From the reported results, it can be concluded that the proposed network is robust to variations in the input image size to a great extent.

3) Effect of Training Image set Size

The dataset originally contains about 28000 images that are unequally distributed amongst the galleries. We select equivalently proportional number of images from each gallery to form a new, smaller size of training image set for each fold of this experiment. Table 9 reports the exact proportion of training set used in our experiments. The proportion is relevant to the original number of images in each image set i.e., for an image set containing 100 images, 70% training image set size would mean 70 images are being used. It is clear from the results that the impact of decreasing the training image sets is significantly low on our model. Even when using only 10% of the original training data, it out performs the state-of-the-art accuracy.

TABLE 8. Effect of Image resolution on accuracy

S. No.	Image Size	MV	WV
1	48×48	100%	100%
2	56×56	100%	100%
3	64×64	100%	99.31%
4	72×72	94.55%	94.55%

TABLE 9. Effect of Training Image set Size

S. No.	Training Image Set Size	MV	WV
1	90%	99.31%	99.31%
2	80%	97.95%	99.31%
3	70%	95.23%	96.59%
4	60%	99.31%	99.31%
5	50%	97.27%	96.59%
6	20%	90.47%	93.87%
7	10%	80.95%	78.23%

V. CONCLUSION

We propose a novel deep learning based image set classification methodology for facial emotion recognition. To the best of our knowledge, previously, the works directed in this field have used single-image or video based inputs. We show, with detailed experimentation and analysis, how the proposed image set classification can improve the accuracy and efficiency of this process. Our work outperforms the state-of-the-art accuracy on FER2013, CK+ and FERG datasets, in addition to achieving superior results for the SFEW dataset. One of the major hurdles we observed in the process are scarcity of images collected in natural settings. Hence, future works can be dedicated for the collection of more in-the-wild facial data.

SHARJEEL TAHIR has completed Research Masters with Training (RMT) at Murdoch University, Perth, WA, Australia. After which, he has worked a Research Associate at the University of Western Australia. He is currently a PhD Candidate at Edith Cowan University, WA, Australia. He received his Bachelor of Computer Science majoring in Robotics from Lahore Garrison University (LGU), Lahore, Punjab, Pakistan. He has participated in and won many national level Robotics competitions. He has also partaken in many individual as well as group level robotics and automation projects. His current research interests include robot vision, deep learning, emotion recognition, empathy in artificial intelligence, and scene understanding.

NIMA MIRNATEGHI has completed Master of Information Technology in Data Science at Murdoch University, Perth, WA, Australia. He received his Bachelor of Computer Science from the University of Wollongong (UOW), NSW, Australia. He is currently a PhD Candidate at Edith Cowan University, WA, Australia. He has participated in a number of IEEE WA competitions, and software/game development hackathons. Prior to starting his degree at Murdoch, he worked as a programming instructor at multiple international schools in Dubai, UAE. His current research interests include deep learning, pattern recognition, statistical analysis, object recognition, facial detection, machine learning, and image processing.

SYED AFAQ ALI SHAH received the PhD degree in computer vision and machine learning from The University of Western Australia (UWA), Crawley, WA, Australia. He was a Lecturer ICT with Central Queensland University, Australia. He is currently a Senior Lecturer at Edith Cowan University, Perth, WA, Australia. He is also an Adjunct Senior Lecturer with the Department of Computer Science and Software Engineering, UWA, Perth, WA, Australia. His current research interests include deep learning, object/face recognition, Scene understanding, and image processing. Dr. Shah was a recipient of the Start Something Prize for Research Impact through Enterprise for 3-D Facial Analysis Project funded by the Australian Research Council. He has authored over 50 research articles and co-authored a book, A guide to convolutional neural networks for computer vision.

FERDOUS SOHEL received a PhD degree from Monash University, Australia. He is currently an Associate Professor in Information Technology at Murdoch University, Australia. He worked as a Research Assistant Professor/Research Fellow at the School of Computer Science and Software Engineering, The University of Western Australia. His research interests include computer vision, machine learning, pattern recognition, and digital agriculture. He is a recipient of prestigious Discovery Early Career Research Award funded by the Australian Research Council. He is an Associate Editor of IEEE Transactions on Multimedia and IEEE Signal Processing Letters. He is a member of the Australian Computer Society and a senior member of the IEEE.

REFERENCES

[1] Michel F. Valstar, Bihan Jiang, Marc Mehu, Maja Pantic, and Klaus Scherer. The first facial expression recognition and analysis challenge. In *2011 IEEE International Conference on Automatic Face Gesture Recognition (FG)*, pages 921–926, 2011.

[2] Latest trends in emotion recognition methods: case study on emotiw challenge. *International Journal of Advanced Computer Research*, 10(46):34–50, 01 2020.

[3] Byoung Chul Ko. A brief review of facial emotion recognition based on visual information. *sensors*, 18(2):401, 2018.

[4] Yin Fan, Xiangju Lu, Dian Li, and Yuanliu Liu. Video-based emotion recognition using cnn-rnn and c3d hybrid networks. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 445–450, 2016.

[5] Zhiding Yu and Cha Zhang. Image based static facial expression recognition with multiple deep network learning. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, page 435–442, New York, NY, USA, 2015. Association for Computing Machinery.

[6] Liqian Liang, Congyan Lang, Yidong Li, Songhe Feng, and Jian Zhao. Fine-grained facial expression recognition in the wild. *IEEE Transactions on Information Forensics and Security*, 16:482–494, 2021.

[7] Syed A. A. Shah, Uzair Nadeem, Mohammed Bennamoun, Ferdous Sohel, and Roberto Togneri. Efficient image set classification using linear regression based image reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

[8] Munawar Hayat, Mohammed Bennamoun, and Senjian An. Deep reconstruction models for image set classification. *IEEE transactions on pattern analysis and machine intelligence*, 37(4):713–727, 2014.

[9] Sharjeel Tahir. Emotenet: Deep neural network for facial emotion recognition using image set classification, 2021.

[10] Ian J. Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, Yingbo Zhou, Chetan Ramaiah, Fangxiang Feng, Ruifan Li, Xiaojie Wang, Dimitris Athanasakis, John Shawe-Taylor, Maxim Milakov, John Park, Radu Ionescu, Marius Popescu, Cristian Grozea, James Bergstra, Jingjing Xie, Lukasz Romaszko, Bing Xu, Zhang Chuang, and Yoshua Bengio. Challenges in representation learning: A report on three machine learning contests, 2013.

[11] Deepali Aneja, Alex Colburn, Gary Faigin, Linda Shapiro, and Barbara Mones. Modeling stylized character expressions via deep learning. In Shang-Hong Lai, Vincent Lepetit, Ko Nishino, and Yoichi Sato, editors, *Computer Vision – ACCV 2016*, pages 136–153, Cham, 2017. Springer International Publishing.

[12] 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, 2011.

[13] Abhinav Dhall, Roland Goecke, Jyoti Joshi, Michael Wagner, and Tom Gedeon. Emotion recognition in the wild challenge 2013. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, ICMI '13, page 509–516, New York, NY, USA, 2013. Association for Computing Machinery.

[14] Hoai-Duy Le, Guee-Sang Lee, Soo-Hyung Kim, Seungwon Kim, and Hyung-Jeong Yang. Multi-label multimodal emotion recognition with transformer-based fusion and emotion-level representation learning. *IEEE Access*, 11:14742–14751, 2023.

[15] Chenghao Zhang and Lei Xue. Autoencoder with emotion embedding for speech emotion recognition. *IEEE Access*, 9:51231–51241, 2021.

[16] Zhang Kexin and Liu Yunxiang. Speech emotion recognition based on transfer emotion-discriminative features subspace learning. *IEEE Access*, 11:56336–56343, 2023.

[17] Chirag Dalvi, Manish Rathod, Shruti Patil, Shilpa Gite, and Ketan Kotecha. A survey of ai-based facial emotion recognition: Features, ml dl techniques, age-wise datasets and future directions. *IEEE Access*, 9:165806–165840, 2021.

[18] Ruhul Amin Khalil, Edward Jones, Mohammad Inayatullah Babar, Tariqullah Jan, Mohammad Haseeb Zafar, and Thamer Alhussain. Speech emotion recognition using deep learning techniques: A review. *IEEE Access*, 7:117327–117345, 2019.

[19] Shichuan Du, Yong Tao, and Aleix M Martinez. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 111(15):E1454–E1462, 2014.

[20] T. M. W. Vithanawasam and B. G. D. A. Madhusanka. Face and upper-body emotion recognition using service robot's eyes in a domestic environment. In *2019 International Research Conference on Smart Computing and Systems Engineering (SCSE)*, pages 44–50, 2019.

[21] Krithika L.B and Lakshmi Priya GG. Student emotion recognition system (sers) for e-learning improvement based on learner concentration metric. *Procedia Computer Science*, 85:767–776, 2016. International Conference on Computational Modelling and Security (CMS 2016).

[22] Maryam Imani and Gholam Ali Montazer. A survey of emotion recognition methods with emphasis on e-learning environments. *Journal of Network and Computer Applications*, 147:102423, 2019.

[23] Barnabás Takács and Harry Wechsler. A dynamic and multiresolution model of visual attention and its application to facial landmark detection. *Computer Vision and Image Understanding*, 70(1):63–73, 1998.

[24] Seyedehsamaneh Shojaeilangari, Wei-Yun Yau, Karthik Nandakumar, Jun Li, and Eam Khwang Teoh. Robust representation and recognition of facial emotions using extreme sparse learning. *IEEE Transactions on Image Processing*, 24(7):2140–2152, 2015.

[25] Kartika Candra Kirana, Slamet Wibawanto, and Heru Wahyu Herwanto. Facial emotion recognition based on viola-jones algorithm in the learning environment. In *2018 International Seminar on Application for Technology of Information and Communication*, pages 406–410, 2018.

[26] Yuanyuan Ding, Qin Zhao, Baoqing Li, and Xiaobing Yuan. Facial expression recognition from image sequence based on lbp and taylor expansion. *IEEE Access*, 5:19409–19419, 2017.

[27] Jung-Wei Hong, Meng-Ju Han, Kai-Tai Song, and Fuh-Yu Chang. A fast learning algorithm for robotic emotion recognition. In *2007 International Symposium on Computational Intelligence in Robotics and Automation*, pages 25–30, 2007.

[28] Javier Ruiz del Solar, Patricio Loncomilla, and Naiomi Soto. A survey on deep learning methods for robot vision, 2018.

[29] Robert Walecki, Ognjen, Rudovic, Vladimir Pavlovic, Björn Schuller, and Maja Pantic. Deep structured learning for facial action unit intensity estimation, 2017.

[30] Heechul Jung, Sihaeng Lee, Junho Yim, Sunjeong Park, and Junmo Kim. Joint fine-tuning in deep neural networks for facial expression recognition. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2983–2991, 2015.

[31] Chuanhe Liu, Tianhao Tang, Kui Lv, and Minghao Wang. Multi-feature based emotion recognition for video clips. New York, NY, USA, 2018. Association for Computing Machinery.

[32] MICHAEL P. PERRONE and LEON N. COOPER. *When networks disagree: Ensemble methods for hybrid neural networks*, pages 342–358.

[33] Zhengning Wang, Fanwei Zeng, Shuaicheng Liu, and Bing Zeng. Oaenet: Oriented attention ensemble for accurate facial expression recognition. *Pattern Recognition*, 112:107694, 2021.

[34] Ling Zhou, Qirong Mao, Xiaohua Huang, Feifei Zhang, and Zhihong Zhang. Feature refinement: An expression-specific feature learning and fusion method for micro-expression recognition. *Pattern Recognition*, 122:108275, 2022.

[35] Mengyi Liu, Shaoxin Li, Shiguang Shan, and Xilin Chen. Au-aware deep networks for facial expression recognition. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–6, 2013.

[36] D Y Liliana. Emotion recognition from facial expression using deep convolutional neural network. *Journal of Physics: Conference Series*, 1193:012004, apr 2019.

[37] Behzad Bozorgtabar, Dwarikanath Mahapatra, and Jean-Philippe Thiran. Exprada: Adversarial domain adaptation for facial expression analysis. *Pattern Recognition*, 100:107111, 2020.

[38] Weijie Sheng and Xinde Li. Multi-task learning for gait-based identity recognition and emotion recognition using attention enhanced temporal graph convolutional network. *Pattern Recognition*, 114:107868, 2021.

[39] Berat A. Erol, Abhijit Majumdar, Patrick Benavidez, Paul Rad, Kim-Kwang Raymond Choo, and Mo Jamshidi. Toward artificial emotional intelligence for cooperative social human–machine interaction. *IEEE Transactions on Computational Social Systems*, 7(1):234–246, 2020.

[40] Yixin Wang, Shuang Qiu, Xuelin Ma, and Huiguang He. A prototype-based spd matrix network for domain adaptation eeg emotion recognition. *Pattern Recognition*, 110:107626, 2021.

[41] Kuang-Chih Lee, J. Ho, Ming-Hsuan Yang, and D. Kriegman. Video-based face recognition using probabilistic appearance manifolds. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 1, pages I–I, 2003.

[42] T.-K. Kim, O. Arandjelovic, and R. Cipolla. Learning over sets using boosted manifold principal angles (bompa). In *Proc. BMVC*, pages 58.1–58.10, 2005. doi:10.5244/C.19.58.

[43] O. Yamaguchi, K.. Fukui, and K.-i. Maeda. Face recognition using temporal image sequence. In *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, pages 318–323, 1998.

[44] Tae-Kyun Kim, Josef Kittler, and Roberto Cipolla. Discriminative learning and recognition of image set classes using canonical correlations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1005–1018, 2007.

[45] Hakan Cevikalp and Bill Triggs. Face recognition based on image sets. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2567–2573, 2010.

[46] Ajmal Mian, Yiqun Hu, Richard Hartley, and Robyn Owens. Image set based face recognition using self-regularized non-negative coding and adaptive distance metric learning. *IEEE Transactions on Image Processing*, 22(12):5252–5262, 2013.

[47] Syed Afaq Ali Shah, Mohammed Bennamoun, and Farid Boussaid. Iterative deep learning for image set based face and object recognition. *Neurocomputing*, 174:866–874, 2016.

[48] Syed A. A. Shah, Uzair Nadeem, Mohammed Bennamoun, Ferdous Sohel, and Roberto Togneri. Efficient image set classification using linear regression based image reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

[49] Rui Wang, Xiao-Jun Wu, and Josef Kittler. Symnet: A simple symmetric positive definite manifold deep learning method for image set classification. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[50] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In Mark W. Jones Xianghua Xie and Gary K. L. Tam, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 41.1–41.12. BMVA Press, September 2015.

[51] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[52] Masaki Nakada, Han Wang, and Demetri Terzopoulos. Acfr: Active face recognition using convolutional neural networks. pages 35–40, 07 2017.

[53] 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, 2020.

[54] Xiaofeng Liu, BVK Vijaya Kumar, Jane You, and Ping Jia. Adaptive deep metric learning for identity-aware facial expression recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 20–29, 2017.

[55] Hui Ding, Shaohua Kevin Zhou, and Rama Chellappa. Facenet2expnet: Regularizing a deep face recognition net for expression recognition. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 118–126. IEEE, 2017.

[56] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2852–2861, 2017.

[57] 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, 2017.

[58] Yousif Khaireddin and Zhuofa Chen. Facial emotion recognition: State of the art performance on fer2013, 2021.

[59] Shervin Minaee and Amirali Abdolrashidi. Deep-emotion: Facial expression recognition using attentional convolutional network, 2019.