IEEE Open Journal of the Computer Society

Received 30 May 2024; revised 16 July 2024; accepted 5 August 2024. Date of publication 14 August 2024; date of current version 3 September 2024. The review of this article was arranged by Associate Editor J. Guo.

Digital Object Identifier 10.1109/OJCS.2024.3443511

InFER++: Real-World Indian Facial Expression Dataset

SYED SAMEEN AHMAD RIZVI, ARYAN SETH, JAGAT SESH CHALLA, AND PRATIK NARANG (Senior Member, IEEE)

Department of Computer Science and Information Systems, Birla Institute of Science and Technology, Pilani, Rajasthan 333031, India

CORRESPONDING AUTHOR: PRATIK NARANG (e-mail: pratik.narang@pilani.bits-pilani.ac.in).

(Syed Sameen Ahmad Rizvi and Aryan Seth contributed equally to this work.)

This work was supported by Kwikpic AI Solutions.

ABSTRACT Detecting facial expressions is a challenging task in the field of computer vision. Several datasets and algorithms have been proposed over the past two decades; however, deploying them in real-world, in-the-wild scenarios hampers the overall performance. This is because the training data does not completely represent socio-cultural and ethnic diversity; the majority of the datasets consist of American and Caucasian populations. On the contrary, in a diverse and heterogeneous population distribution like the Indian subcontinent, the need for a significantly large enough dataset representing all the ethnic groups is even more critical. To address this, we present InFER++, an India-specific, multi-ethnic, real-world, in-the-wild facial expression dataset consisting of seven basic expressions. To the best of our knowledge, this is the largest India-specific facial expression dataset. Our cross-dataset analysis of RAF-DB vs InFER++ shows that models trained on RAF-DB were not generalizable to ethnic datasets like InFER++. This is because the facial expressions change with respect to ethnic and socio-cultural factors. We also present LiteXpressionNet, a lightweight deep facial expression network that outperforms many existing lightweight models with considerably fewer FLOPs and parameters. The proposed model is inspired by MobileViTv2 architecture, which utilizes GhostNetv2 blocks to increase parametrization while reducing latency and FLOP requirements. The model is trained with a novel objective function that combines early learning regularization and symmetric cross-entropy loss to mitigate human uncertainties and annotation bias in most real-world facial expression datasets.

INDEX TERMS Facial expression recognition, annotation bias mitigation, affective computing.

I. INTRODUCTION

Human facial expressions serve as a highly potent and prominent signal for communicating emotional states and intentions [1], [2]. Exhaustive studies [3] have established that among multiple channels conveying human emotions, such as voice, text, gestures, and facial expressions, the latter is regarded as the most precise. Mehrabian et al. [4] confirmed that facial expressions account for 55% of the communication of feelings, while spoken words constitute 7%, and paralinguistic elements make up the remaining 38%. Consequently, facial expressions have been proven to play a critical role in the process of information exchange. Ekman et al. [5], [6], based on a cross-cultural study, identified six fundamental emotions as universal markers by which humans recognize basic emotions across different cultures. These core emotions are anger, fear, disgust, happiness, sadness, and surprise. Over the past two decades, the rapid advancements in the design and development of artificial intelligence systems have led to numerous efforts in automatic facial affect recognition. Due to their significance in Human-Computer Interaction (HCI), facial expressions have gained increasing importance in robotics, healthcare, behavioural research, smart surveillance, and various other HCI applications.

Facial Expression Recognition (FER) systems can be classified into two primary categories: traditional methods and deep learning-based methods. In the early stages, traditional handcrafted methods were developed, utilizing feature extraction techniques such as local binary patterns, non-negative

matrix factorization, and Gabor filters. With the advancements in deep learning and artificial intelligence, a multitude of approaches have emerged, utilizing deep learning paradigms such as CNNs [7], [8], [9], GANs [10], deep belief networks [11], vision transformers [12], [13], [14] and recurrent neural networks [15].

Deep learning extracts high-level, complex abstractions through layers of hierarchical non-linear transformations. Being a supervised learning approach, deep FER models implicitly require adequately large training data representative of various scenarios. For a FER system to effectively cater to a diversified population, it is crucial to have a training data that reflects the diversity in the target group. This is because facial expressions can vary based on factors such as age, gender, ethnicity, and other socio-cultural aspects. Therefore, the diversity in the dataset is the key to the scalability of an FER system. Moreover, the majority of the FER datasets consist of captured posed expressions under controlled lab scenarios; they lack in the wild real-world data consisting of different backgrounds, occlusions, non-frontal head poses, and poor lighting conditions. A deep FER system trained on such limited data often results in suboptimal performance and cannot scale well to real-world applications.

The Indian subcontinent, renowned for its rich socio-cultural diversity, encompasses a vast array of ethnic groups. A study [16] examining the ethnic diversity of the Indian population identifies several ethnic groups, including Dravidian, Mongoloid, Mongolo-Dravidian, Aryo-Dravidian, Indo-Aryan, Scytho-Dravidian, and Turko-Iranian. Thus, implementing an effective, real-world FER system for such an assorted population presents a significant challenge without a dataset that sufficiently represents all ethnic and demographic groups. Therefore, for this highly diverse population, there is a crucial need for a comprehensive dataset that can address the varied requirements of the Indian subcontinent, home to over 1.4 billion people.

A. OUR CONTRIBUTIONS

In this paper, we introduce InFER++, a diversified India-specific crowdsourced real-world facial expression dataset designed to encompass a wide range of ethnicities and demographics across the Indian subcontinent. We also propose LiteXpressionNet, a lightweight deep FER model for facial expression classification. The significant contributions of this paper include:

- The InFER++ dataset comprises 35,789 crowdsourced images representing seven fundamental expressions–anger, disgust, fear, happiness, neutral, sadness, and surprise, spanning various Indian ethnicities, demographic groups, and genders, along with their gender and age labels. To the best of our knowledge, this is the largest India-specific multi-ethnic facial expression dataset.
- We present a cross-dataset analysis between RAF-DB vs InFER++ to demonstrate that existing real-world FER datasets do not generalize well to ethnic-specific datasets. Therefore, substantiating the need for a real-world ethnically diverse dataset.
- We present a comprehensive performance analysis of the recent open-source state-of-the-art (SOTA) deep FER methods on the proposed dataset.
- LiteXpressionNet, a novel lightweight deep FER model. We show a comprehensive experimentation analysis on RAF-DB, CK+, and AffectNet, datasets with respect to various open-source SOTA lightweight deep FER methods.

II. RELATED WORK

Facial expressions are among the most powerful, instinctive, and universal indicators of human emotional states and intentions. Within the fields of computer vision and machine learning, extensive research has been conducted on facial expression recognition (FER) systems to capture and analyze expression data from facial images. This section reviews some recent and significant advancements in the FER domain.

A. FER DATASETS

A thorough dataset with sufficient labelled training data that encompasses a broad spectrum of variations in terms of ethnicities, age groups, genders, and environments is essential for the development of a robust deep FER system. In this section, we delve into some of the existing FER datasets that are available in the literature. Table 1 presents a summary of the most commonly referenced FER datasets in the literature.

CK+ [17] is widely recognized as the most frequently utilized lab-controlled dataset for assessing FER systems. The dataset includes 593 video clips recorded from 123 subjects. These sequences, spanning 10 to 60 frames, illustrate the change from a neutral to a peak facial expression. JAFFE [18] dataset is a curated collection of 213 posed facial expression images from 10 Japanese women created under controlled lab settings. Each participant is depicted in 3 to 4 images for each of the seven fundamental facial expressions (anger, disgust, fear, happiness, sadness, and surprise). The dataset possesses a challenge because of the sparse number of examples for each subject and expression. FER2013 [20] dataset, posed as a challenge during ICML-2013, is a crowdsourced, unconstrained dataset automatically collected using the Google image search API. The database contains 28,709 training, 3,589 validation, and 3,589 test images, each annotated with one of seven basic expressions. MMI [19] dataset comprises 740 image sequences and 2900 videos from 25 subjects, each displaying seven basic expressions. The frame sequences in this dataset initiate with a neutral expression, attain peak expression intensity mid-sequence, and subsequently return to a neutral expression.

The CMU Multi-Pie [22] dataset consists of 755,370 frames, capturing six facial expressions from 337 subjects. These images are recorded from 15 distinct viewpoints and under 19 varying illuminating conditions across four different recording sessions. The Oulu-CASIA [27] dataset comprises 2,880 images sourced from 80 subjects. Videos were recorded

TABLE 1. A Summary of Current Facial Expression Datasets. ['-' Denotes Not Applicable.]

Dataset	Samples		# Subject	Condition		Elicitation		Expression Distribution
	# Images	# Videos		Lab	Scraped	Posed	Spontaneous	
CK+ [17]	5,876	593	123	✓	✗	✓	✓	7 basic emotions + contempt
JAFFE [18]	213	-	10	✓	✗	✓	✗	7 basic emotions
MMI [19]	740	2900	25	✓	✗	✓	✗	7 basic emotions
FER-2013 [20]	35,887	-	-	✗	✓	✓	✓	7 basic emotions
TFD [21]	112,234	-	-	✓	✗	✓	✗	7 basic emotions
Multi-PIE [22]	755,370	-	337	✓	✗	✓	✗	disgust, neutral, surprise, smile, squint, and scream.
AFEW 7.0 [23]	1,809	-	-	✗	✓	✓	✓	7 basic emotions
SFEW 2.0 [24]	1,766	-	-	✗	✓	✓	✓	7 basic emotions
BU-3DFE [25]	2500 (3D)	-	100	✓	✗	✓	✗	7 basic emotions
ExpW [26]	91,793	-	-	✗	✓	✓	✓	7 basic emotions
Oulu-CASIA [27]	2,880	-	80	✓	✗	✓	✗	6 basic emotions without neutral
EmotioNet [28]	1,000,000	-	-	✗	✓	✓	✓	23 basic + compound emotions
ISED [29]		428	50	✓	✗	✗	✓	disgust, happiness, sadness, and surprise.
RAF-DB [8]	29,762	-	-	✗	✓	✓	✓	7 basic + 12 compound emotions
AffectNet [30]	450,000	-	-	✗	✓	✓	✓	7 basic emotions
InFER [31]	10,200	4200	600	✓	✓	✓	✓	7 basic emotions
InFER++	35,789	-	-	✗	✓	✓	✓	7 basic emotions

under two lighting conditions: near-infrared (NIR) and visible light (VIS). The sequences begin with a neutral expression and peak at the final three frames, corresponding to videos captured by the Visible Light system (VIS) under various indoor illumination conditions. The Expression in the Wild (ExpW) [26] dataset is a compilation of facial images retrieved using the Google search API. It contains 91,793 manually labelled faces categorized into seven basic expression classes.

Indian spontaneous expression (ISED) [29] dataset contains 428 videos of 50 samples, consisting of 29 males and 21 females. The captured emotions included happiness, surprise, disgust, and sadness. RAF-DB [8] comprises of 15,339 images depicting various facial expressions, sourced from the internet and manually annotated through crowdsourcing and reliable estimation. This dataset includes seven basic expressions along with eleven compound expressions.

The Indian Facial Expression Recognition (InFER) dataset [31] was our initial attempt to create an India-specific dataset for facial expression recognition. This dataset includes seven basic expressions from 600 individuals, covering a wide range of Indian ethnicities, age groups, and genders, along with their corresponding labels. In total, the dataset comprises 10,200 images and 4,200 video clips. Of these, 4,200 images capture the seven basic expressions from the 600 subjects, while the remaining 6,000 images were crowd-sourced from the web to ensure diversity and comprehensiveness.

The InFER++ dataset, to the best of our knowledge, is the largest India-specific facial expression dataset, comprising 35,789 images representing seven basic facial expressions. Unlike most FER datasets, InFER++ also includes gender and age labels, which are essential for investigating and mitigating bias in facial expression recognition. Fig. 1 provides a visual comparison of InFER++ versus the most popular previously existing FER datasets. Fig. 1(b) provides some samples of occlusions in the Indian context which include turbans, ornaments, and facial hairs. Some outlier scenarios such as poor illumination conditions, and non-frontal head poses are also depicted.

B. FACIAL EXPRESSION RECOGNITION METHODS

The field of FER has garnered considerable focus in recent years, with an increasing number of researchers concentrating on the area of human-computer interaction. In the initial years, various studies have explored traditional and handcrafted feature extraction methods such as local binary patterns [32], local binary patterns on three orthogonal planes [33], non-negative matrix factorization [34], gabor wavelets [35], and sparse learning [36] for FER.

With the evolution of deep learning and various FER challenges, including FER-2013, and EmotiW etc, many Convolutional Neural Network (CNN) based FER methods were proposed. Savchenko et al. [7] explored the use of EfficientNet [37] for FER tasks and demonstrated the efficacy of convolutional neural networks for this purpose. Sang et al. [38] concentrated on minimizing the variation within the same class of facial expression depth features using a dense convolutional network for the task of facial expression recognition. Zhang et al. [39] suggested a method of erasing attention consistency to tackle the task of FER with noisy labelled data.

As vision transformers (ViT) based models gained attention in the recent past, a variety of facial expression recognition methods that incorporate transformers have been proposed in the literature. The first application of the vision transformer to the study of FER was by Xue et al. [40] which resulted in state-of-the-art performance. VTFF [14] stands out in handling facial expression recognition tasks in uncontrolled environments because of its feature fusion capabilities. Zhao et al. [41] integrated both global and local attention to tackle

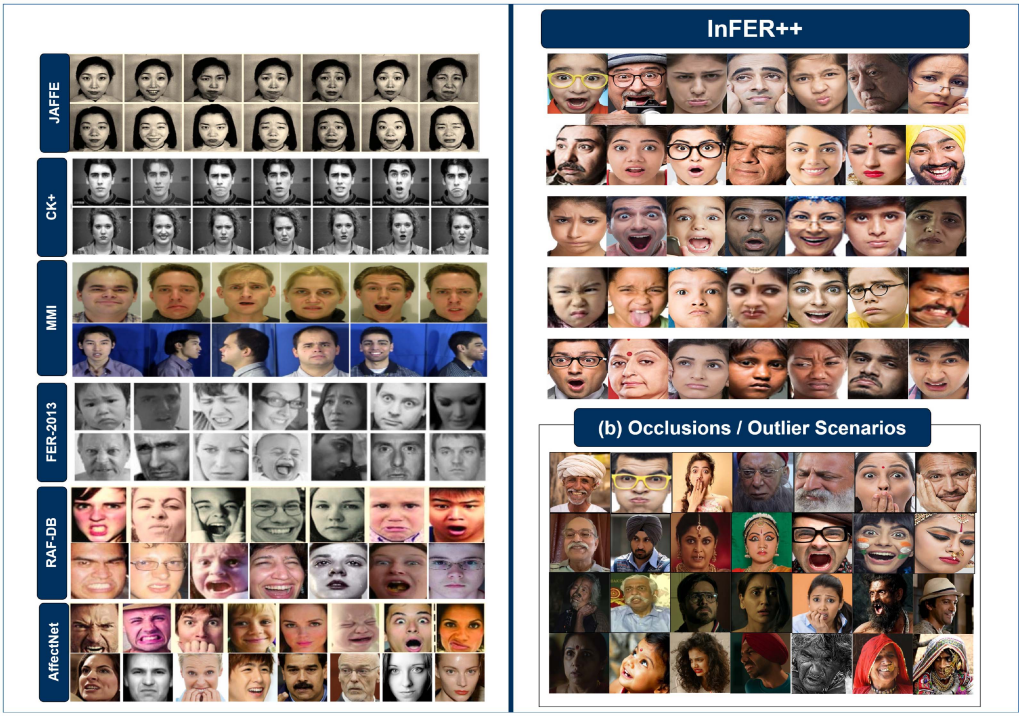

FIGURE 1. A visual comparison between InFER++ and previous FER datasets (JAFFE, CK+, MMI FER-2013, RAF-DB, and AffectNet). 1(b) represents some samples consisting of occlusions and other outlier scenarios such as poor illumination conditions, and non-frontal head poses. [Best viewed in colour].

the significant challenges of occlusion and pose variation in facial expression recognition. POSTER [42] addresses the issues of intra-class discrepancy, inter-class similarity, and scale sensitivity in Facial Expression Recognition (FER) simultaneously. It achieves this by merging image features with facial landmark features using a two-stream, cross-fusion, and pyramid design approach. Zhao et al. proposed Efficient-Face [43], a robust and accurate facial expression recognition (FER) network with fewer parameters, designed for FER in the wild. To enhance robustness, the network includes a local feature extractor and a channel-spatial modulator using depthwise convolution, enabling awareness of both local and global salient facial features. Additionally, the training strategy incorporates a novel and efficient label distribution learning (LDL) method to account for the combinations and mixtures of basic emotions. In [44], the authors designed a facial expression recognition (FER) system aimed at supporting healthcare applications. To this end, they proposed various convolutional neural network (CNN) architectures capable of handling multiresolution facial images. In [45], authors propose a residual-based CNN architecture, termed BReG-Net. It comprised 39 layers of residual blocks, in which the traditional skip-connections in the ResNet blocks were substituted with a differentiable function featuring a bounded gradient. This modification effectively mitigates the issues of gradient vanishing and exploding. In continuation to their previous work [49], the authors introduced BReG-NeXt in [46], which incorporates a greater number of trainable parameters and provides enhanced flexibility. In FaceTopoNet [47], authors propose a deep FER model that learns an effective tree topology of the face. FaceTopoNet features two streams: one for learning facial structure, focusing on landmark positions, and another for learning texture, focusing on patches around the landmarks. The outputs of these streams are fused using an attention-based fusion strategy. In [48], authors propose a novel deep FER network that utilizes representational learning based on latent spaces to mitigate bias in FER systems.

III. THE InFER++ DATASET

The facial expressions vary significantly across different ethnicities, age groups and genders. The lack of an ethnically diverse dataset is a significant bottleneck whilst applying FER models to real-world wild practical applications. The proposed InFER++ is our attempt to enhance a diverse India-specific dataset encompassing all the Indian ethnicities, age groups, and genders covering real-world in-the-wild scenarios. InFER++ is a continuation of InFER [31], our initial attempt to develop a diverse India-specific FER dataset. The proposed InFER++, as far as we know, is the most extensive India-specific dataset for real-world expressions. This section describes the methodology adopted in the creation of the dataset.

A. DATA COLLECTION

An image search API was used to extract URLs using a specifically curated dictionary of search queries. These queries were curated in a manner that search results yielded images that have a diverse ethnic (in the Indian context) demographic

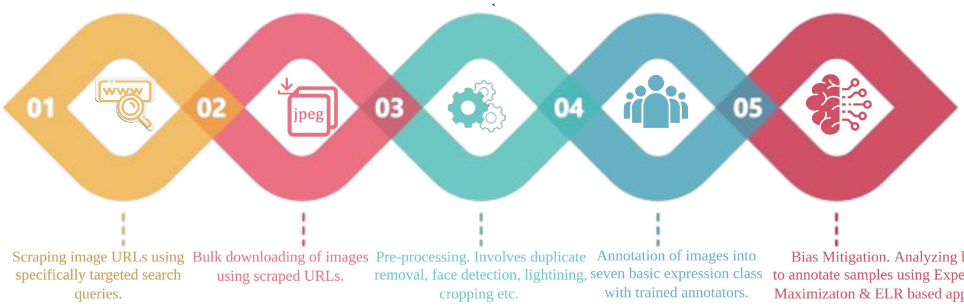

FIGURE 2. **Overview of data collection methodology of InFER++.**

composition. The dictionary had 9 segments using which search queries were formulated; these included six segments of basic expressions, i.e. anger, disgust, fear, happy, sad and surprise. The dictionary had synonyms of 9 segments using which search queries were formulated; these included six segments of basic expressions, i.e. anger, disgust, fear, happy, sad and surprise. The remaining segments corresponded to age, gender and ethnicity. Some examples of these search queries included 'angry punjabi child', 'disgusted adolescent gujrati girl', 'fearful teenager tamil boy', 'happy elderly telugu female', 'surprised kashmiri man', 'excited assami boy', 'depressed bengali male', 'annoyed marathi female' etc. Here the keywords - punjabi, marathi gujrati, Kashmiri, assami, tamil, telugu and bengali represent the rich ethnic composition of the Indian subcontinent.

India, being a diverse socio-cultural nation, has about 22 different languages, each belonging to a specific cultural and ethnic heritage. To encompass this diverse ethnicity, some of the most common search queries were translated into languages like Marathi, Bengali, Tamil, Telugu, Malayali, Kannada, Punjabi, Meitei and Urdu. This approach helped in scraping some of the ethnically images from the World Wide Web. Images were then downloaded using extracted URLs yielded by the image search API. We then ran a RetinaFace [49] face detection module to filter out the images without human faces. Consequently, about 35789 facial images were obtained from a total of 50,371 images. Fig. 2 provides an overview of the data collection methodlogy of InFER++. Fig. 3 shows some sample images in InFER++.

B. DATASET ANNOTATION

Annotating 35,000 facial images accurately to their respective expression, gender and age class requires an in-depth psychological understanding of emotional states. To tackle this, a group of 50 annotators were trained in the demographic and psychological knowledge of seven basic expressions. Post-training, an assessment was conducted on a sample dataset, which consisted of 1500 images from a few lab-based as well real world datasets. On the basis of their performance, the best 30 annotators were selected to annotate our scraped images. Images were equally and randomly allotted to each annotator, ensuring that there was no direct correlation between images labelled by an individual. It was ensured that each image was labelled by at least 10 annotators. In the end, a multi-label annotation result in the form of a 7-dimensional vector is obtained, where each dimension correlates to respective votes for one of the seven basic expressions for each image.

C. MITIGATING ANNOTATION BIAS

Expression classification requires niche and expert psychological understanding of emotional states and the way they are conveyed by facial expressions. Labelling the data with a group of annotators without the aforementioned expertise incorporates the individual biases in the dataset. Moreover, the subjectivity and wide-ranging difficulty of labels further complicate the annotation process. As a result, disagreements were obvious between different labellers.

To mitigate this problem of noisy labels, we adopt a strategy inspired by [8], [50], which used the expectation maximization (EM) approach followed by using the Early Learning Regularization (ELR) based framework to determine the incorrectly labelled images. Note that these images do not necessarily need to be incorrectly labelled but were re-annotated by a group of five expert annotators from the core team.

Let $D = \{(x_j, y_j, t^1, t^2 \ldots, t^R)\}_{j=1}^{n}$ indicate a set of n annotated inputs where y_i is the actual label (hidden variable) for jth samples x_j, $t_j^i \in 1, 2, 3, 4, 5, 6, 7$ is the corresponding label given by the ith annotator in the superscript The corresponding probabilities of t_j^i is expressed as a sigmoid function.

$$p(t_j^i = y_j | \alpha^i, \beta_j) = \frac{1}{1 + e^{-\alpha^i . \beta_j}} \tag{1}$$

where $1/\beta_j$ is the difficulty of jth sample, α^i is the reliability of ith annotator.

The objective here is to optimize the log-likelihood of the given labels.

$$\max_{\beta>0} l(\alpha, \beta) = \sum_j \ln p(t \mid \alpha, \beta) = \sum_j \ln \sum_y p(t, y \mid \alpha, \beta)$$
$$= \sum_j \ln \sum_y Q_j(y) \frac{p(t, y \mid \alpha, \beta)}{Q_j(y)}$$

FIGURE 3. Sample images of seven basic expressions in InFER++ post pre-processing and Annotation. [Best viewed in colour].

$$\geq \sum_{j} \sum_{y} Q_j(y) \ln \frac{p(t, y \mid \alpha, \beta)}{Q_j(y)} \tag{2}$$

where $Q_j(y)$ is a certain distribution of hidden variable y,

$$Q_j\left(y_j\right) = \frac{p\left(t_j, y_j \mid \alpha, \beta\right)}{\sum_y p\left(t_j, y_j \mid \alpha, \beta\right)} = \frac{p\left(t_j, y_j \mid \alpha, \beta\right)}{p\left(t_j \mid \alpha, \beta\right)}$$
$$= p\left(y_j \mid t_j, \alpha, \beta\right) \tag{3}$$

After revision, annotations of 24 labelers remained and with the remaining annotated samples we then proceeded with finding out the extremely hard-to-annotate samples with early learning regularization-based method. About 6,754 hard-to-annotate images were then re-annotated with the core expert annotator team with their respective labels. The detailed algorithm is described in Algorithms 1 & 2. Recognizing the budget constraints, we only selected the k images which had a high loss value. This was done to ensure that the annotators were able to spend enough time per image.

Facial Expression Recognition annotations are afflicted by human visual uncertainty. While this uncertainty has been used to improve classification performance in [51], we established that these images must be more carefully analyzed. A visual description of some of these images has been provided in Fig. 4.

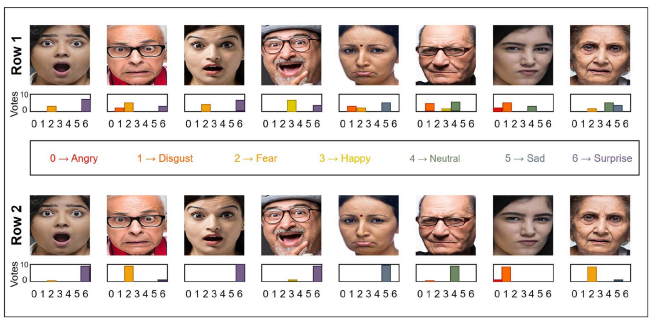

FIGURE 4. Some sample images of hard-to-annotate expressions. Row 1 corresponds to the initial expression distribution as annotated by 30 trained annotators. Row 2 corresponds to expression distribution after the annotation bias mitigation strategy was employed. [Zoomable upto 5X.] .

D. METADATA

The distribution of the 35,789 images across different expressions, age groups and gender in the InFER++ dataset is detailed in Table 2. The dataset also provides the age and gender labels. We divided the age groups into five distinct categories: children (5–10 years), teenagers (11–19 years), young adults (20–29 years), adults (30–59 years), and senior citizens (60 years and above). Of the 35,789 images collected, 2,326 were of children, 7,050 were of teenagers, 10,880 were of young adults, 12,061 were of adults, and 3,472 were of senior citizens. There was about an equal proportion of both

Algorithm 1: Annotation Reliability Estimation.

Data: Training set $D = (x_j, t_j^1, t_j^2, \ldots, t_j^R,)_{j=1}^n$

Result: Each annotator's reliability α_i^*

1: **Initialize:** $\forall j = 1, \ldots, n$, initialize the true label y_j using majority voting

$$\beta_j := -\sum_{i=1}^{R} p\left(t_j^i\right) \ln p\left(t_j^i\right), \alpha_i := 1$$

The initial value of β_j is image j's entropy. The greater the entropy the more uncertain the image.

2: **Repeat:** E-Step:

$$Q_j\left(y_j\right) := \prod_i p\left(y_j \mid t_j, \alpha_i, \beta_j\right)$$

M-step:

$$\alpha_i := \arg\max_{\alpha_i} \sum_j \sum_{y_j} Q_j\left(y_j\right) \ln \frac{p\left(t_j, y_j \mid \alpha_i, \beta_j\right)}{Q_j\left(y_j\right)}$$

Until convergence

Algorithm 2: Algorithm for Estimating Noisiness of Label.

Input: Dataset $\mathcal{D}$ with n images

Result: Ranked noisiness of labels

Initialize: criterion ← Modified-ELRLoss()

rank ← {[-1, i] | $i \in (1, n)$} ▷ Set all entries in rank array to -1 model ← ResNet18()

Train(model, $\mathcal{D}$) ▷ Train ResNet18 on dataset $\mathcal{D}$ criterion2 ← CrossEntropyLoss, defined as:

$$L(y, \hat{y}) = -\sum_{c=1}^{C} y_c \log(\hat{y}_c)$$

for $i \leftarrow 1$ **to** n **do**

    output ← model($\mathcal{D}[i]$.image)

    loss ← criterion(output, $\mathcal{D}[i]$.label)

    rank[i][0] ← loss

sort(rank) ▷ Select top k based on budget

genders, with 18,650 images corresponding to males, whereas 17,139 images belonged to the female class.

E. RAF-DB VS INFER++ CROSS DATASET STUDY

Subsequently, we also performed a cross-dataset study of RAF-DB vs InFER++ to investigate the differences in expression classification with respect to a diverse multi-ethnic set of populations. In this context, a cross-dataset study implies that we utilize all the images from one dataset for training purposes and the images from different dataset for testing. We then performed a cross-database experiment based on six-class expression. Multiclass classification using a lightweight vision transformer (MobileVIT) and confusion matrix was used as the classification method and the assessment criteria, respectively. Fig. 5 shows the results of this experiment.

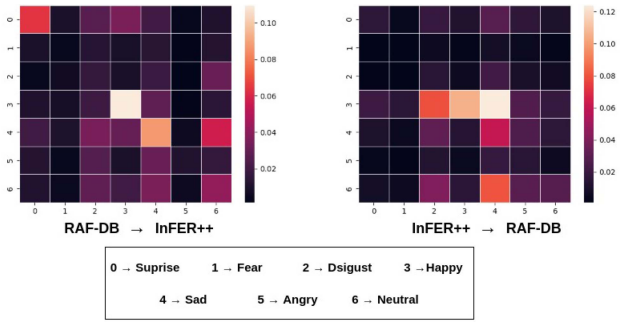

FIGURE 5. Heatmaps of confusion matrix of MobileNet for cross-dataset study. (Best viewed in colour.).

Upon analyzing the diagonals of the confusion matrix, it can be inferred that surprise, happy, and disgust are the three expressions with the highest classification accuracies in both scenarios. Furthermore, the average accuracy of the RAF-DB → InFER analysis turned out to be 33%, whereas that of the InFER→ RAF-DB analysis was just 23%. This substantiates the need for having a diverse multi-ethnic dataset. Expression classes like sadness and fear, anger and disgust, and happiness and surprise were often confused and misclassified among each other, which is also one of the reasons for a lower performance on this cross-dataset study. This cross-study strongly demonstrates how existing State-of-the-Art datasets do not generalize well to ethnicity-specific data and how ethnicity-specific data by itself cannot generalize to other ethnicities. This demonstrates a significant lack of ethnic variance in SOTA open-source datasets.

F. PERFOMANCE ANALYSIS OF INFER++

We also carried out experiments to analyse the performance of the proposed dataset on recent open source state-of-the-art deep FER methods. This section provides a detailed description of the methods and the experimental results achieved thereafter. The images were cropped using RetinaFace to the size of 256 × 256 pixels and converted into grayscale for further classification. The images were partitioned into training and test set, with 80% images of each expression being used training and the rest of the 20% for testing purposes. All the experiments were carried out on an Intel (R) Xeon 2.20GHz CPU with NVIDIA A5000 with 24GB of graphical memory.

1) DEEP FER METHODS

In recent years, several deep FER methods have been presented. Below, we discuss some of the latest SOTA open-source methods utilized in our experiments.

EfficientNet FER: Savchenko et al. [7] developed EfficientNet for predicting student engagement and emotions in online learning environments and MobileNet for real-time video analysis on mobile devices.

TABLE 2. Expression, Age and Gender Distribution of InFER++. (a) Represents the Number of Images Per Expression, (b) Demonstrates Images Per Age Group, and (c) Shows the Gender Distribution of InFER++ Dataset

(a)

Expression	# Images
Anger	3,799
Disgust	1,025
Fear	3,855
Happy	13,078
Neutral	7,851
Sad	4,391
Surprise	1,790

(b)

Age Group	# Images
Children (5-10)	2,326
Teenagers (11-19)	7,050
Young Adults (20-29)	10,880
Adults (30-59)	12,061
Senior Citizens (60+)	3,472

(c)

Gender	# Images
Males	18,650
Females	17,139

TABLE 3. Performance Analysis of InFER++ on SOTA Open Source Deep FER Methods

Model	Method	Anger	Disgust	Fear	Happy	Neutral	Sad	Surprise	Average Accuracy (%)
Savchenko *et al.* [7]	EfficientNet	0.8510	0.836	0.8610	0.921	0.881	0.901	0.852	87.11
SCN [53]	Suppressing Uncertainties	0.866	0.849	0.880	0.931	0.892	0.916	0.852	88.1
EfficientFace [44]	**Label Distribution Learning**	**0.874**	**0.858**	**0.886**	**0.941**	**0.903**	**0.926**	**0.867**	**89.21**
DACL [9]	ResNet18	0.869	0.853	0.878	0.941	0.894	0.912	0.850	88.39
Ad-Corre [55]	Ensemble Learning	0.836	0.813	0.841	0.901	0.864	0.871	0.842	85.34
Island Loss [56]	Island Loss	0.821	0.806	0.831	0.891	0.854	0.861	0.840	84.34
Howard *et al.* [57]	MobileNetV3	0.782	0.767	0.792	0.861	0.816	0.826	0.810	80.76

SCN: Wang et al. [52] introduced the Self Cure Network (SCN) to address inconsistent FER annotations with a self-attention mechanism, ranking regularization, and relabeling strategies.

EfficientFace: Zhao et al. [43] proposed a lightweight, robust FER network using a local-feature extractor, channel-spatial modulator, and depthwise convolution to learn global facial features with an efficient learning strategy.

DACL: DACL [9] employs an attention mechanism to enhance feature discrimination, using spatial feature maps to estimate attention weights and sparse center loss formulation for better intra-class compactness and inter-class separation.

Ad-Corre: Fard et al. [53] introduced Adaptive Correlation Loss to improve FER by learning feature vectors with high correlation within the same class and low correlation between different classes.

Island Loss: Cie et al. [54] proposed an island loss-based CNN to reduce intra-class variation and enlarge inter-class variations, enhancing the discriminating power of deep features. They also discussed a center loss-based CNN approach used in experimentation.

MobileNetV3: Optimized for mobile CPUs, MobileNetV3 [55] combines network architecture search (NAS) with the NetAdapt algorithm and architectural advancements for improved performance on mobile devices.

2) PERFORMANCE ANALYSIS

We run each model for 50 epochs from scratch using the training configuration to the best of our ability from code provided by the authors or through training configurations mentioned in the paper. For cases where neither are clearly available, we run Adam Optimizer [56] at a learning rate of 0.001 with $\beta_1 = 0.99$ and $\beta_2 = 0.999$.

Table 3 lists the deep FER methods along with their corresponding accuracies on our dataset. From the results in Table 3 it can be inferred that our proposed dataset performs reasonably well on the SOTA deep FER methods. We also observe that EfficientFace [43] yields the highest accuracy among all the methods, achieving an average accuracy of about 89.21%. We also observe that among the seven expression classes, happy, sad and neutral have comparatively better accuracies: disgust, anger, fear and surprise.

IV. LITEXPRESSIONNET

We propose a Lightweight Architecture for Facial Expression Recognition. Facial Expression Recognition as a field is moving towards mobilizing and deploying models in the wild. For deployment, these models must be able to be deployed on the edge and have low latency. We base our model around the MobileViTv2 [57] architectures which combine CNNs and Attention modules. We further utilize GhostNetV2 [58] blocks to increase parametrization while reducing latency and FLOPS requirements. This approach comes with the advantage of utilizing parameters effectively. Motivated by Neural Architecture Search literature [59] where it is shown that model accuracy is highly correlated to the number of parameters in a model, our design philosophy focuses on increasing the number of parameters while keeping the FLOPS for our model constant. Fig. 6 shows the network architecture for LiteXpressionNet. We use the MBConv blocks, which use depthwise convolutions to capture the spatial interaction. This choice is made because of the FFN modules in MBConv and Attentional Modules.

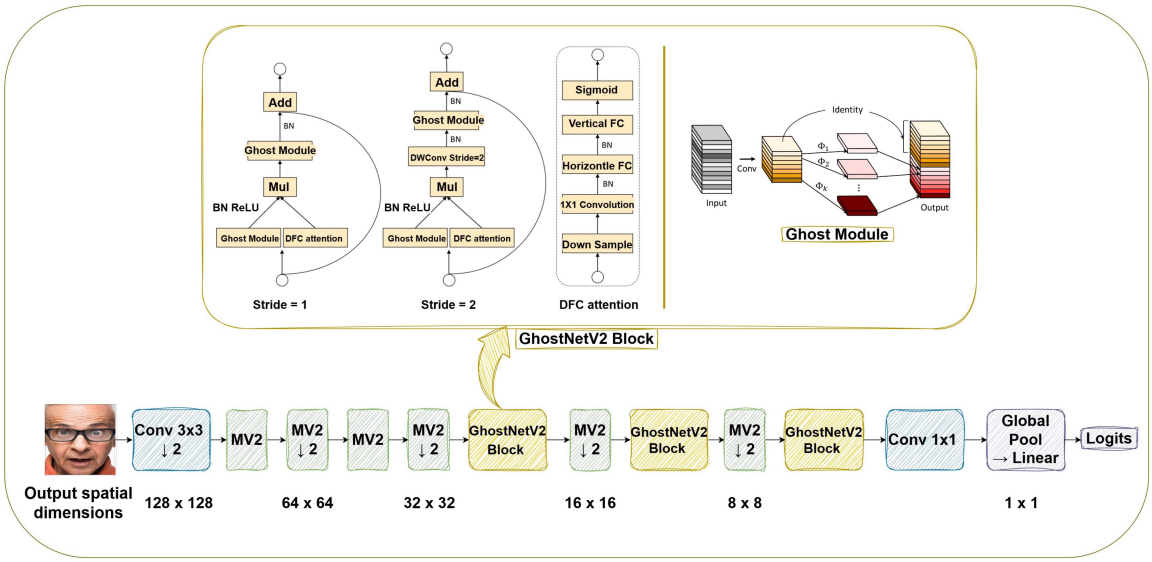

FIGURE 6. Network architecture of LiteXpressionNet. [Zoomable upto 5X].

To decrease latency, we replace the MobileViT V2 blocks with GhostNetv2 blocks. This was done to include as many parameters as possible without increasing latency. It is optimal since the GhostNet block incorporates global and local features with a low-latency DFC attention module.

Objective Function: Facial Expression Recognition has to deal with annotation bias and large amounts of human uncertainty for certain images. Mitigating this, we propose a loss function based on Early Learning Regularization(ELR) [60] and Symmetric Cross Entropy loss to limit the effect of noisy labels on overall accuracy. Our loss function is defined by:

$$\mathcal{L}_{\mathrm{ELR}}(\Theta) := \mathcal{L}_{\mathrm{SCE}}(\Theta) + \frac{\lambda}{n}\sum_{i=1}^{n}\log\left(1 - \left\langle \mathbf{p}^{[i]}, \mathbf{t}^{[i]}\right\rangle\right)$$

This method alternates between computing the target outputs and minimizing the loss via gradient descent. The target estimation is done using the following equation:

$$\mathbf{t}^{[i]}(k) := \beta\mathbf{t}^{[i]}(k-1) + (1-\beta)\mathbf{p}^{[i]}(k)$$

Here β is the momentum term, $t^{[i]}$ is the target value for the ith example and $p^{[i]}$ denotes the model output for the ith example.

A. EXPERIMENTATION, RESULTS AND ANALYSIS

To evaluate the effectiveness of the proposed method, experiments were conducted on three widely used FER datasets. For this purpose we select CK+ [17] one of the most popular lab-based FER dataset, along with RAF-DB [8] and AffectNet [30] two most widely used in-the-wild real-world FER datasets. We assess the seven basic emotions in the CK+ dataset using a 10-fold cross-validation method to ensure a fair comparison with the majority of existing research. For RAF-DB, we evaluate seven basic expressions, which consist of about 12,271 training samples and 3,068 test samples. Similarly, for AffectNet, we evaluate our proposed method for 8 expression classes consisting of about 287,568 training samples and 4,000 test samples.

All the experiments were carried out on an Intel (R) Xeon 2.20 GHz CPU with NVIDIA A5000 with 24 GB of graphical memory. We compare the accuracy with FLOPS and demonstrate how LiteXpressionNet is optimal for deployment on edge devices. LiteXpressionNet was trained using an Adam Optimizer [56] with Cosine Annealing [69] and Sharpness-Aware Minimization.

From Tables 4 and 5 it can inferred that LiteXpressionNet achieves comparable performance on fewer FLOPS. This allows for better deployment on edge devices and is a more comprehensive model. Further, we note that EfficientFace [43] requires an additional component during training, which means it cannot be trained or fine-tuned on resource-constrained systems.

Furthermore, analyzing Table 4 we can conclude that on RAF-DB, LiteXpressionNet outperforms all the rest of the methods, barring EfficientFace. It also stands as the lightest model with the least number of parameters, and FLOPS, with the exception of EfficientFace, which has a lesser number of parameters. Similarly, for AffectNet 8 class classification task, the model outperforms the rest of the methods, with the exception of EfficentFace and SCN. For lab-based dataset, analysing the performance on CK+ dataset, with 10-fold cross-validation, we observe that LiteXpressionNet outperforms all the other lightweight deep FER methods.

V. CONCLUSION

Facial expression recognition (FER) has garnered significant interest due to its crucial applications in social robotics, healthcare, driver fatigue monitoring, and human-computer interaction systems. Despite extensive research, models trained on existing datasets often underperform in real-world

TABLE 4. Comparisons of SOTA Lightweight FER Methods on In-the-Wild Datasets. RAF-DB (7 Basic Expressions) and AffectNet (8 Expressions)

Dataset	Method	# Flops (M)	# Parameters (M)	Accuracy(%)
RAF-DB	EfficientNet [7]	380	5.3	86.12
	MobileNetv3-Large [57]	225	5.4	78.12
	SCN [53]	1818.56	11.8	87.03
	gACNN [63]	15479.79	134.29	85.07
	ShuffleNetV2 [64]	2330	20.31	84.31
	Xception [65]	4630	22.86	85.29
	EdgeNeXt-S [66]	5600	5.6	84.66
	RAN [67]	14548.45	11.19	86.90
	LDL-ALSG [68]	4109.48	23.52	85.53
	FGENet [69]	2050	28.2	86.72
	EfficientFace [44]	154.18	**1.2**	**88.36**
	LiteXpressionNet(Ours)	**126**	4.6	86.34
AffectNet-8	VGGNet-Variant [70]	80.44	6.54	58.00
	MobileNet-Variant [70]	13.56	0.074	56.00
	SCN [53]	1818.56	11.18	**60.23**
	EfficientFace [44]	154.18	**1.28**	59.89
	LiteXpressionNet(Ours)	**126**	4.6	58.56

TABLE 5. Comparisons of SOTA Lightweight FER Methods on Lab-Based Dataset (CK+)

Method	Accuracy(%)
IACNN [72]	95.37
DLP-CNN [8]	95.78
gACNN [63]	96.40
SCAN-CCI [73]	97.31
DeRL [74]	97.37
FN2EN [75]	98.60
ShuffleNetV2 [64]	91.83
Xception [65]	91.82
EdgeNeXt-S [66]	90.25
FGENet [69]	97.89
LiteXpressionNet (Ours)	**98.92**

scenarios due to insufficient socio-cultural and ethnic diversity, with a dominance of American and Caucasian subjects. Addressing this gap, this paper introduces InFER++, a diverse, India-specific FER dataset comprising 35,789 images representing seven basic expressions, annotated with gender and age labels.

The images were collected via a curated dictionary of search queries using an image search API and annotated by 30 trained annotators, ensuring each image received at least 10 annotations. A bias mitigation strategy, combining expectation maximization and early learning regularization, was employed to enhance labeling accuracy. Cross-dataset analysis between RAF-DB and InFER++ demonstrates the limited generalizability of existing datasets to ethnically diverse populations.

We also present a performance evaluation of recent state-of-the-art (SOTA) deep FER methods on InFER++ and introduce LiteXpressionNet, a novel lightweight FER model. LiteXpressionNet achieves comparable accuracy to SOTA models while significantly reducing the number of parameters and FLOPS. Future research will focus on mitigating bias and enhancing FER model robustness by curating datasets targeting various protected attributes. InFER++ and LiteXpressionNet are valuable tools for developing and deploying FER systems in diverse, real-world applications.

REFERENCES

[1] C. Darwin and P. Prodger, *The Expression of the Emotions in Man and Animals*. Oxford, U.K.: Oxford Univ. Press, 1998.

[2] Y.-I. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 2, pp. 97–115, Feb. 2001.

[3] P. Ekman, "*Emotions revealed: Recognizing Faces and Feelings to Improve Communication and Emotional Life*. New York, NY, USA: Times books, 2003.

[4] A. Mehrabian and J. A. Russell, *An Approach to Environmental Psychology.* Cambridge, MA, USA: The MIT Press, 1974.

[5] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion.," *J. Pers. Soc. Psychol.*, vol. 17, no. 2, 1971, Art. no. 124.

[6] P. Ekman, "Strong evidence for universals in facial expressions: A reply to russell's mistaken critique," *Psychol. Bull.*, vol. 115, pp. 268–287, 1994.

[7] A. V. Savchenko, L. V. Savchenko, and I. Makarov, "Classifying emotions and engagement in online learning based on a single facial expression recognition neural network," *IEEE Trans. Affect. Comput.*, vol. 13, no. 4, pp. 2132–2143, Oct.–Dec. 2022.

[8] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2852–2861.

[9] A. H. Farzaneh and X. Qi, "Facial expression recognition in the wild via deep attentive center loss," in *Proc. IEEE/CVF winter Conf. Appl. Comput. Vis.*, 2021, pp. 2402–2411.

[10] H. Yang, Z. Zhang, and L. Yin, "Identity-adaptive facial expression recognition through expression regeneration using conditional generative adversarial networks," in *2018 13th IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2018, pp. 294–301.

[11] P. Liu, S. Han, Z. Meng, and Y. Tong, "Facial expression recognition via a boosted deep belief network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1805–1812.

[12] F. Xue, Q. Wang, Z. Tan, Z. Ma, and G. Guo, "Vision transformer with attentive pooling for robust facial expression recognition," *IEEE Trans. Affect. Comput.*, vol. 14, no. 4, pp. 3244–3256, Oct.–Dec. 2023.

[13] A. Chaudhari, C. Bhatt, A. Krishna, and P. L. Mazzeo, "Vitfer: Facial emotion recognition with vision transformers," *Appl. System Innov.*, vol. 5, no. 4, 2022, Art. no. 80.

[14] F. Ma, B. Sun, and S. Li, "Facial expression recognition with visual transformers and attentional selective fusion," *IEEE Trans. Affect. Comput.*, vol. 14, no. 2, pp. 1236–1248, Apr.–Jun. 2023.

[15] N. Jain, S. Kumar, A. Kumar, P. Shamsolmoali, and M. Zareapoor, "Hybrid deep neural networks for face emotion recognition," *Pattern Recognit. Lett.*, vol. 115, pp. 101–106, 2018.

[16] H. H. Risley, "The study of ethnology in India," *J. Anthropological Inst. Great Britain Ireland*, vol. 20, pp. 235–263, 1891.

[17] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (CK+) : A complete dataset for action unit and emotion-specified expression," in *2010 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.-Workshops.*, 2010, pp. 94–101.

[18] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with gabor wavelets," in *Proc. 3rd IEEE Int. Conf. Autom. Face Gesture Recognit.*, 1998, pp. 200–205.

[19] M. Valstar et al., "Induced disgust, happiness and surprise: An addition to the mmi facial expression database," in *Proc. 3rd Intern. Workshop EMOTION (Satellite LREC): Corpora Res. Emotion Affect*. Paris, France., 2010, Art. no. 65.

[20] I. J. Goodfellow et al., "Challenges in representation learning: A report on three machine learning contests," in *Proc. Int. Conf. Neural Inf. Process.*, 2013, pp. 117–124.

[21] J. M. Susskind, "Interpreting faces with neurally inspired generative models," Ph.D. dissertation, Univ. Toronto, 2011, pp. 64–78.

[22] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," *Image Vis. Comput.*, vol. 28, no. 5, pp. 807–813, 2010.

[23] A. Dhall, R. Goecke, S. Ghosh, J. Joshi, J. Hoey, and T. Gedeon, "From individual to group-level emotion recognition: Emotiw 5.0," in *Proc. 19th ACM Int. Conf. Multimodal Interact.*, 2017, pp. 524–528.

[24] A. Dhall, O. Ramana Murthy, R. Goecke, J. Joshi, and T. Gedeon, "Video and image based emotion recognition challenges in the wild: Emotiw 2015," in *Proc. 2015 ACM Int. Conf. Multimodal Interact.*, 2015, pp. 423–426.

[25] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3D facial expression database for facial behavior research," in *Proc. IEEE 7th Int. Conf. Autom. face gesture Recognit.*, 2006, pp. 211–216.

[26] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "From facial expression recognition to interpersonal relation prediction," *Int. J. Comput. Vis.*, vol. 126, no. 5, pp. 550–569, 2018.

[27] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. PietikäInen, "Facial expression recognition from near-infrared videos," *Image Vis. Comput.*, vol. 29, no. 9, pp. 607–619, 2011.

[28] C. Fabian Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, "Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5562–5570.

[29] S. Happy, P. Patnaik, A. Routray, and R. Guha, "The Indian spontaneous expression database for emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 8, no. 1, pp. 131–142, Jan.–Mar. 2017.

[30] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 18–31, Jan.–Mar. 2019.

[31] S. Rizvi, P. Agrawal, J. Challa, and P. Narang, "InFER: A multi-ethnic Indian facial expression recognition dataset," in *Proc. 15th Int. Conf. Agents Artif. Intell.- Vol. 3:*, SciTePress, 2023, pp. 550–557.

[32] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image Vis. Comput.*, vol. 27, no. 6, pp. 803–816, 2009.

[33] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 915–928, Jun. 2007.

[34] R. Zhi, M. Flierl, Q. Ruan, and W. B. Kleijn, "Graph-preserving sparse nonnegative matrix factorization with application to facial expression recognition," *IEEE Trans. Syst., Man, Cybern., Part B. (Cybern.)*, vol. 41, no. 1, pp. 38–52, Feb. 2011.

[35] H.-B. Deng et al., "A new facial expression recognition method based on local gabor filter bank and PCA plus LDA," *Int. J. Inf. Technol.*, vol. 11, no. 11, pp. 86–96, 2005.

[36] L. Zhong, Q. Liu, P. Yang, J. Huang, and D. N. Metaxas, "Learning multiscale active facial patches for expression analysis," *IEEE Trans. Cybern.*, vol. 45, no. 8, pp. 1499–1510, Aug. 2015.

[37] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[38] D. V. Sang, L. T. B. Cuong and P. T. Ha, "Discriminative deep feature learning for facial emotion recognition," in *2018 IEEE 1st Int. Conf. Multimedia Anal. Pattern Recognit.*, 2018, pp. 1–6.

[39] Y. Zhang, C. Wang, X. Ling, and W. Deng, "Learn from all: Erasing attention consistency for noisy label facial expression recognition," in *Eur. Conf. Comput. Vis.*, 2022, pp. 418–434.

[40] F. Xue, Q. Wang, and G. Guo, "Transfer: Learning relation-aware facial expression representations with transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 3601–3610.

[41] Z. Zhao, Q. Liu, and S. Wang, "Learning deep global multi-scale and local attention features for facial expression recognition in the wild," *IEEE Trans. Image Process.*, vol. 30, pp. 6544–6556, 2021.

[42] C. Zheng, M. Mendieta, and C. Chen, "Poster: A pyramid cross-fusion transformer network for facial expression recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 3146–3155.

[43] Z. Zhao, Q. Liu, and F. Zhou, "Robust lightweight facial expression recognition network with label distribution training," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 4, pp. 3510–3519 .

[44] C. Bisogni, A. Castiglione, S. Hossain, F. Narducci, and S. Umer, "Impact of deep learning approaches on facial expression recognition in healthcare industries," *IEEE Trans. Ind. Inform.*, vol. 18, no. 8, pp. 5619–5627, Aug. 2022.

[45] B. Hasani, P. S. Negi, and M. H. Mahoor, "Bounded residual gradient networks (BREG-NeT) for facial affect computing," in *2019 14th IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2019, pp. 1–7.

[46] B. Hasani, P. S. Negi, and M. H. Mahoor, "BReG-NeXt: Facial affect computing using adaptive residual networks with bounded gradient," *IEEE Trans. Affect. Comput.*, vol. 13, no. 2, pp. 1023–1036, Apr.–Jun. 2022.

[47] M. Kolahdouzi, A. Sepas-Moghaddam, and A. Etemad, "FaceTopoNet: Facial expression recognition using face topology learning," *IEEE Trans. Artif. Intell.*, vol. 4, no. 6, pp. 1526–1539, Dec. 2023.

[48] S. S. A. Rizvi, A. Seth, and P. Narang, "FAIR-FER: A latent alignment approach for mitigating bias in facial expression recognition (student abstract)," in *Proc. AAAI Conf. Artif. Intell.*, 2024, vol. 38, no. 21, pp. 23633–23634.

[49] S. Serengil and A. Ozpinar, "A benchmark of facial recognition pipelines and co-usability performances of modules," *Bilisim Teknolojileri Dergisi*, vol. 17, no. 2, pp. 95–107, 2024. [Online]. Available: https://dergipark.org.tr/en/pub/gazibtd/issue/84331/1399077

[50] J. Whitehill, P. Ruvolo, T. Wu, J. Bergsma, and J. Movellan, "Whose vote should count more: Optimal integration of labels from labelers of unknown expertise," in *Proc. 22nd Int. Conf. Neural Inf. Process. Syst.*, 2009, pp. 2035–2043.

[51] J. C. Peterson, R. M. Battleday, T. L. Griffiths, and O. Russakovsky, "Human uncertainty makes classification more robust," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9617–9626.

[52] K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao, "Suppressing uncertainties for large-scale facial expression recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6897–6906.

[53] A. P. Fard and M. H. Mahoor, "Ad-Corre: Adaptive correlation-based loss for facial expression recognition in the wild," *IEEE Access*, vol. 10, pp. 26756–26768, 2022.

[54] J. Cai, Z. Meng, A. S. Khan, Z. Li, J. O'Reilly, and Y. Tong, "Island loss for learning discriminative features in facial expression recognition," in *2018 IEEE 13th Int. Conf. Autom. Face Gesture Recognit.*, 2018, pp. 302–309.

[55] A. Howard et al., "Searching for MobileNetV3," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 1314–1324.
[56] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[57] S. Mehta and M. Rastegari, "Separable self-attention for mobile vision transformers," *Trans. Mach. Learn. Res.*, vol. 2023, 2023. [Online]. Available: https://openreview.net/forum?id=B4/BEKI
[58] Y. Tang, K. Han, J. Guo, C. Xu, C. Xu, and Y. Wang, "GhostNetV2: Enhance cheap operation with long-range attention," *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 9969–9982, 2022.
[59] T. Yang, L. Yang, X. Jin, and C. Chen, "Revisiting training-free nas metrics: An efficient training-based method," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2023, pp. 4751–4760.
[60] S. Liu, J. Niles-Weed, N. Razavian, and C. Fernandez-Granda, "Early-learning regularization prevents memorization of noisy labels," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 20331–20342, 2020.
[61] Y. Li, Y. Lu, J. Li, and G. Lu, "Separate loss for basic and compound facial expression recognition in the wild," in *Proc. 11th Asian Conf. Mach. Learn.*, 2019, pp. 897–911.
[62] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "ShuffleNet V2: Practical guidelines for efficient CNN architecture design," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 122–138.
[63] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1251–1258.
[64] M. Maaz et al., "Edgenext: Efficiently amalgamated CNN -transformer architecture for mobile vision applications," in *Proc. Eur. Conf. Computer Vis.*, 2022, pp. 3–20.
[65] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, "Region attention networks for pose and occlusion robust facial expression recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 4057–4069, 2020.
[66] S. Chen, J. Wang, Y. Chen, Z. Shi, X. Geng, and Y. Rui, "Label distribution learning on auxiliary label space graphs for facial expression recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 13984–13993.
[67] M. Sun and C. Yan, "FGENet: A lightweight facial expression recognition algorithm based on fasternet," *Signal, Image Video Process.*, vol. 18, pp. 5939–5956, 2024.
[68] C. Hewitt and H. Gunes, "CNN-based facial affect analysis on mobile devices," 2018, *arXiv:1807.08775*.
[69] L. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *Proc. 6th Int. Conf. Learn. Representations*, Toulon, France, Apr. 2017, pp. 24–26. [Online]. Available: https://openreview.net/forum?id=Skq89Scxxx.TIONAL neural network for TCIAL SYDression recompITION*arXiv:1608.03983*.
[70] Z. Meng, P. Liu, J. Cai, S. Han, and Y. Tong, "Identity-aware convolutional neural network for facial expression recognition," in *2017 IEEE 12th Int. Conf. Autom. Face Gesture Recognit.*, 2017, pp. 558–565.
[71] D. Gera and S. Balasubramanian, "Landmark guidance independent spatio-channel attention and complementary context information based facial expression recognition," *Pattern Recognit. Lett.*, vol. 145, pp. 58–66, 2021.
[72] D. Ruan, Y. Yan, S. Lai, Z. Chai, C. Shen, and H. Wang, "Feature decomposition and reconstruction learning for effective facial expression recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 7660–7669.
[73] D. Ruan, Y. Yan, S. Lai, Z. Chai, C. Shen, and H. Wang, "Feature decomposition and reconstruction learning for effective facial expression recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 7660–7669.

SYED SAMEEN AHMAD RIZVI received the B.Tech degree in computer science and engineering from Jamia Hamdard, New Delhi, India, in 2019. He is currently working toward the Ph.D. degree with the Department of Computer Science & Information Systems, Birla Institute of Technology & Science (BITS), Pilani, Rajasthan, India. He is also a Research Scholar with the Department of Computer Science and Information Systems, BITS Pilani, under the supervision of Professor Pratik Narang. His research focuses on leveraging advanced deep learning techniques to address complex challenges in the field of affective computing.

ARYAN SETH is currently working toward the Undergraduate degree (third-year) in computer science with the Birla Institute of Technology & Science, Pilani, Rajasthan, India. His research interests include artificial intelligence, machine learning, and software development and has demonstrated his skills through research initiatives and collaborative ventures. He is also dedicated to leveraging his knowledge to solve real-world problems and is committed to continuous learning and professional growth in the field of computer science.

JAGAT SESH CHALLA received the B.Sc., M.E., and Ph.D. degrees from the Birla Institute of Technology and Science (BITS), Pilani, Rajasthan, India. He was a Postdoctoral Researcher with Northwestern University, Evanston, IL, USA. He is currently an Assistant Professor with the Department of Computer Science and Information Systems, Birla Institute of Technology & Science. His research interests include high-performance data mining, stream analytics, Big Data analytics, HPC, stream analytics, computer vision, and federated learning. He has also founded the Postman API and Coding Innovation Laboratory, BITS Pilani.

PRATIK NARANG (Senior Member, IEEE) received the M.Sc. (Tech.) and Ph.D. degrees from the Birla Institute of Technology and Science (BITS), Pilani, Rajasthan, India. He was on Research positions with the National University of Singapore, Singapore, and New York University Abu Dhabi, Abu Dhabi, UAE. He is currently an Associate Professor of CSIS with BITS Pilani. He is the Faculty Advisor with Google Developer Student Clubs, BITS Pilani. His research focuses on computer vision and NLP using machine/deep learning. He has also led multiple sponsored R&D projects funded by the Indian government and industry and collaborates with top universities (NUS, UIUC) and companies (IBM, NVIDIA, Birdeye) globally.