ERFN: Leveraging Context for Enhanced Emotion Detection

Navneet Gupta[1], R. Vishnu Priya[2], Chandan Kumar Verma[3]
Department of Mathematics, Bioinformatics and Computer Applications, Maulana Azad National Institute of Technology, Bhopal, Madhya Pradesh, India, 462003[1, 3]
Department of Computer Applications, National. Institute of Technology, Tiruchirappalli, Tamil Nadu, India[2]

***Abstract*—The majority of previous methods for identifying emotions concentrate on facial expressions rather than taking into account the rich contextual information that suggests significant emotional states. To fully utilize the contextual information in order to make up for the lack of emotion information. In this work, The Emotion Recognition Fusion Network (ERFN) is a novel model that uses advanced techniques for efficient context-aware identification of human emotion recognition. It incorporates the Flow Context Aware Loss Fusion (FCALF) model, which focuses on emotion analysis in a video sequence. The model uses deep feature extraction (VGG16), Farnebäck optical flow model, and L1 loss to calculate the Average Contextual Loss (ACL) for selecting key frames. The selected frames are used to obtain resultant optical flow images. Data augmentation techniques are applied exclusively to the training images. The resultant optical flow images undergo feature extraction using both InceptionResNetV2 and VGG16, fine-tuned by adding layer followed by GlobalMaxPool2D and a dense layer, capturing intricate details and flow-contextual information from face, body, and scene. The fused features are fed into a Softmax layer for classification. Experimental results show that the ERFN outperforms existing models in terms of accuracy and generalization, contributing to its effectiveness in capturing context-aware emotions. The proposed approach shows promising results in real-world uncontrolled environments (CAER-S) and laboratory-controlled (CK+) datasets.**

Keywords*—Context-based emotion recognitions; deep learning; optical flow; CNN

I. Introduction

Emotion is a fundamental aspect of life with significant impact on human thinking, knowledge, and decision-making. This plays an essential role recently in robotics, healthcare, education, and human-computer interaction[1, 2]. The majority of earlier research has adopted human-derived modalities, includes voice, text and facial emotion. Among all the facial emotion is most recent trend that resulting in a large number of facial emotion datasets and algorithms [3-5]. Earlier, facial emotion analyses are mainly developed using controlled dataset which was collected from the person who are professional actor. Therefore, uncertainty in the dataset. In addition, the dataset generated with constrained environment has uniform illumination, subtle background variation, frontal imaging or no head movement [6], which is quite different from realistic environment.

The context is another component that is evidenced by psychological research. It has a big impact on how people perceive emotions, according to [7-10]. For example, the same facial expressions in different situations might indicate different mental feeling, such as a person laugh heavily at comedy club versus the same person pouring tears at a funeral. Context: the environment, people around, and situational clues play important cures. So, researchers have focused to the important cures significant by contextual information. Hence, in this work the contextual information fuse with facial expression for robust emotional perception.

To extract valuable contextual information which implies important emotion states, we introduce the attention mechanism for three dynamic features: face, body language and environmental information. To reduce the scale difference between the small face portions and the wide contextual background, we specifically identify the features around the body area as local contextual features and the others as context in general. The proposed work developed a novel the Flow Context Aware Loss Fusion (FCALF) Model for emotion recognition, which is based on the non-overlapping face, body, and environmental context components. The model adopts the typical VGG16 and optical flow model, to extract spatial feature and intricate dynamic motion respectively. To preserve the semantic and contextual features, a new contextual loss function is proposed in this work. At first, the model carefully selects context-aware frame pairings from video sequences which is a crucial task in our technique.

VGG16 model is employed to extract contextual information from facial expression, body language and contextual environment for the selected frames in which emotions are seen. Simultaneously, Farneback optical flow is employed to detect the intricate dynamics motion exhibited in the frame. The outcomes of the VGG16 and flow are averaged through contextual loss function to identify four best frame pairings, which effectively capture the essence of dynamic interactions in a context-rich environment. The optical flow enables the encoding of spatial-temporal dynamics that are crucial for identifying emotions. For further enhancement, the model utilizes advanced transfer learning techniques such as InceptionResNetV2 and VGG16.

The aforementioned methodology makes significant contributions to the field of emotion identification in context-rich environments:

- We proposed a new FCALF model for emotion recognition, which assists the emotion by contextually learning the relationship between face, body and context environment.

- We designed VGG16 and Farneback optical flow as a backbone of the model to extracts spatial-temporal dynamic features.
- A new average contextual loss function is proposed on the dynamic features to select best 4 frame pairs. Those selected pairs are rich contextual information which distinguish the importance of each part such as face, body and environment. This selective approach optimizes computational resources while maximizing the relevance of the extracted features for emotion recognition tasks.
- Fine-tuning pre-trained models extracts high-level features, improving the system's robustness and generalization capabilities

II. Related Work

1) Emotion recognition based on face and body: Most of the research that has been done on recognizing emotions in people has focused on faces, thinking that emotions can be inferred from the way people look. A lot of research has been done on face emotion recognition in the last few years [4, 11-15]. Early works mostly used face images taken in controlled laboratories [16], which only showed a few different head poses, lighting conditions, and other things. Recent research [14, 15] looks into how to recognize facial expressions in the wild. The emotions are also natural and come in a variety of forms. For recognizing facial expressions, traditional techniques mostly use hand-crafted appearance and geometry features taken from the whole face or specific local face regions. These include SIFT [11], LBP [11, 16], and PHOG [12]. These features are then fed into supervised classifiers. SVM [17], random forests [13], and others to figure out how people are feeling. Most of the new work is built on deep learning and uses Convolutional Neural Networks (CNNs) to understand feelings and extract facial features [4, 14, 15, 18], and they do excellent work.

Since body language is also a big part of showing emotions [8], some other ways of detecting emotions use things like hand, shoulder, body movements, and so on. Karpouzis et al. [19] use hand moves to get information about emotions. In their study [20], Nicolaou et al. combine cues from shoulder movements and face reactions to figure out how people are feeling. A brain model by Schindler et al. [21] suggests that body language can be used to figure out how someone is feeling. Yang and Narayanan [22] use a model of body language dynamics to figure out how people are feeling when they are interacting with each other. Recently, deep learning has also been looked at for recognizing body language emotions [23-25]. According to Barros et al. [23], a Multichannel CNN can recognize emotions in both the face and the upper body. In [25], Nguyen et al. suggest a new feature-level fusion method based on multimodal dense bilinear pooling to combine different types of emotions cues, such as body language, facial expressions, and poses.

Face-based and body-based emotions recognition systems are limited because they only look at certain parts of the target person's face and body. However, in real life, there are many clues from the image's background that can be used to figure out how someone is feeling, but these programs don't take them into account. The face and body in the center may also be occluded or not visible, which is something that these networks can't really handle.

2) Recognizing emotions in real-life scenarios: An individual's face and body are often seen along with the main scene in real life, which can greatly affect how that person perceives emotions [26], [27]. Lee et al. [10] recently came up with the idea of the Context-Aware Emotion Recognition Networks (CAER-Net) to help computers understand how people feel in real life. In order to take advantage of the scene environments, they hide people's faces in the picture and model their contributions in a way that is similar to and stronger than those of the human face areas. They also made a collection called Context-Aware Emotion Recognition (CAER) that has a lot of TV show video clips that have been labelled with emotion categories. Their suggested method, on the other hand, doesn't carefully model the inputs of different areas, and it can't really handle hidden or invisible faces, which is a common problem in real life. An emotional graph was made by Zhang et al. [28] using environments to help recognize emotions. It is based on the graph convolution network. The background cues, on the other hand, are only used to improve the main body parts and aren't really thought about for recognizing human emotions. The shapes of the main and background cues are not used as much as they could be. Mittal et al. [29] suggest recognizing emotions from many sources, such as the target person's faces and gaits, as well as the background scene. However, the analysis is not thorough enough to model how each of these sources contributes in particular. Some body parts, like body language, are also not taken into account when figuring out how someone is feeling. to include rich contextual information from the face, body, and scene, the Proposed model improves emotion recognition and greatly increases accuracy and generalization. It supports better patient-caregiver relations as well as the diagnosis and monitoring of mental health disorders in the medical field [30]. to adapt instructional strategies to students' emotional responses, education can improve learning outcomes and provide emotional support [31].

Most of the time, these methods worked pretty well, but they had trouble applying to different face emotions and environments. Researchers started looking for ways to add environmental information and temporal changes to emotion recognition systems after realizing that static analysis had its limits.

3) Integration of contextual information: Incorporating contextual information has become one of the most important ways to improve the accuracy and strength of mood detection. Studies by Kosti et al. [7] and Weixin Li et al. [32] showed that adding scene features, body language, and social environments to tasks for recognizing emotions worked well. The Flow

Context Aware Loss Fusion (FCALF) model we suggested in our study builds on this by mixing information about the scene's context with information about how light moves through it. The FCALF model finds the most useful frame pairs in video sequences by finding the Average Contextual Loss of VGG16 [33] and Farneback optical flow [34] features. This improves the representation of emotional expressions.

4) Temporal dynamics and optical flow analysis: Optical flow analysis is a key part of catching time variations and motion patterns in video clips [35]. Simonyan et al. [36] and Tran et al. [37] showed that visual flow can be useful for tasks like recognizing actions and analyzing gestures. When it comes to recognizing emotions, visual flow analysis lets you pull out changing body language and facial expressions, which gives you useful time information for figuring out what someone is feeling. Our study uses the benefits of both static face features and dynamic motion cues to make emotion detection better. It does this by mixing optical flow analysis with deep transfer learning methods.

5) Deep transfer learning for feature extraction: Deep learning techniques have revolutionized the domain of computer vision by enabling pre-trained models to leverage their acquired knowledge for performing specialized tasks inside their domain. Transfer learning in emotion detection involves extracting high-level features from optical flow images. This facilitates the display of various types of emotions. The experiments conducted by [38] and Zhang et al. [39] demonstrated the effectiveness of deep transfer learning in context-aware emotion detection tasks. These investigations revealed that deep transfer learning is capable of capturing spatial and temporal variations in emotional responses across various environments.

III. PROPOSED WORK

We introduce a sophisticated and efficient Emotion Recognition Fusion Network (ERFN) as seen in Fig. 2 for the purpose of emotion recognition. Our main objectives are: (1) replacing conventional image-level facial features with ERFN, which integrates cutting-edge techniques for context-aware emotion detection. The model incorporates the novel Flow Context Aware Loss Fusion (FCALF) model, as depicted in Fig. 1. This model combines deep feature extraction, L1-loss, and optical flow to compute the Average Contextual Loss (ACL) value. It then identifies the top 4 pairs of frames with the highest ACL value for improved spatial-temporal analysis. Subsequently, the chosen key pairs of frames are used to acquire optical flow images. (2) the two resulting optical flow images are used as input for feature extraction using fine-tuned pretrained models, specifically InceptionResNetV2 [40] and VGG16 [33]. Data augmentation is exclusively incorporated during training, contributing to improved model generalization across diverse emotion recognition scenarios. (3) The output of the two pre-trained models is concatenated and fed into the Softmax for classification. The detail description about each step is discussed as follows:

A. Flow Context Aware Loss Fusion (FCALF)

In the real-time video analysis of Context aware emotion recognition, emphasizing face, body, and scene components, a flow context aware loss fusion model is applied to strategically select the most informative frames. Employing VGG16 and Farnebäck optical flow technique, the algorithm focuses on the face, body, and scene region to capture subtle changes in expressions. Concurrently, body language is analysed through pose estimation, expanding the understanding of emotions to encompass gestures and posture. The broader environmental context is considered, with scene analysis providing insights into contextual elements. This algorithm systematically selects the best 4 key pairs of frames based on the highest ACL value for enhanced spatial-temporal analysis. The selected key pairs of frames are then used to obtain optical flow images by evaluating criteria such as facial expression coverage, comprehensive body language representation, and scene richness, ensuring a concise yet comprehensive representation of emotional cues. The integration of these components in the selected frames through fusion mechanisms within a FCALF model (see Fig. 1) enhances the interpretative depth, offering subtle insights into the emotional states of individuals in real-world scenarios.

Suppose that given a dataset D consists of $VO_i = \{VO_1, VO_2, \dots, VO_k\} \in \mathbb{R}^{\mathrm{D \times N}}$, N samples from M different classes, where$1 \geq \mathrm{i} \leq \mathrm{k}$ and k is the total number of videos. Each video in VO_i is chosen to extract the frames. For subsequent processing, the collected frames of VO_i are saved in a local directory as shown

$$\mathrm{I}_i = \{\mathrm{I}_0, \mathrm{I}_1 \dots \dots, \mathrm{I}_{n-1}\} \quad (1)$$

The pre-trained VGG16 model is utilized for feature extraction from the reference frame $(\mathrm{I}_0)$and subsequent frames (I_t) $where,\ t = 1,2, \dots n-1$ in the image sequence. The layers up to the 23rd layer of the VGG16 model is selected for feature extraction. Each frame in the image sequence is preprocessed using a set of transformations. The preprocessing includes resizing the frames to (224, 224) pixels and converting them to tensor format.

The feature extract from the Reference frame (I_0) can be denoted by I_{fr_0} and the feature extract from the subsequent frames $\{\mathrm{I}_1, \mathrm{I}_2 \dots \dots, \mathrm{I}_t\}$ can be denoted by $\{\mathrm{I}_{fr\_1}, \mathrm{I}_{fr\_2} \dots \dots, \mathrm{I}_{fr\_t}\}$ $where,\ t = 1,2, \dots n-1$.

Take first frame as a Reference frame I_0 and Current frame I_1. Optical Flow $\mathcal{F}_t = (u, v)$ $where\ t = 1,2, \dots n-1$ represents the displacement of pixel (x, y) in the reference frame to its corresponding position in the current frame.

Let's derive the Farnebäck optical flow energy function. The goal is to minimize the energy function with respect to the motion vectors (u, v). The energy function is given by:

$$\mathcal{F}_t = \sum_{x,y}\left((\mathrm{I}_0(x, y) - \mathrm{I}_t(x + u, y + v))\right)^2 . \mathcal{G}\left(\left||x, y|\right|^2; \sigma\right) (2)$$

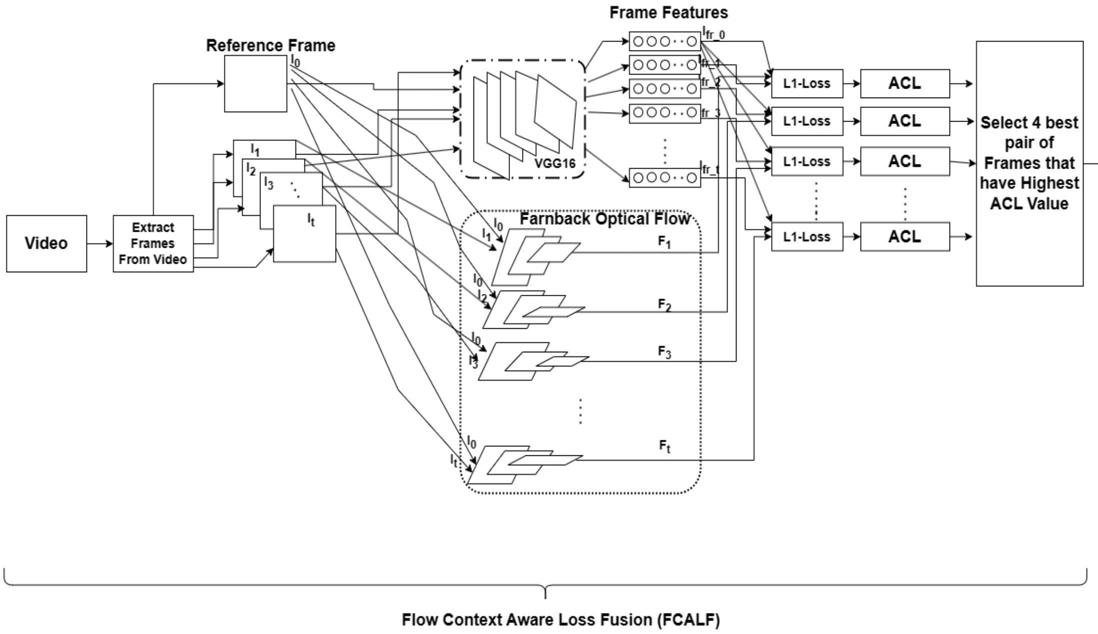

Fig. 1. FCALF Model Architecture: Deep Features and Optical Flow Integration

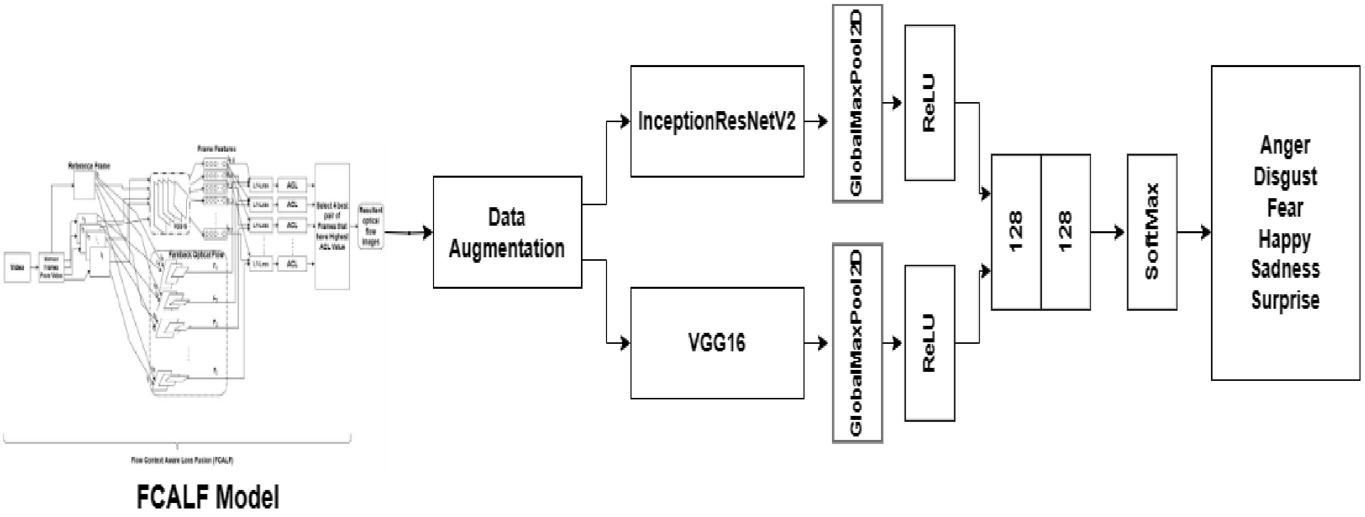

Fig. 2. The diagram illustrates the architectural structure of the Emotion Recognition Fusion Network (ERFN) model

$\mathcal{F}_t$ is the energy function to be minimized, representing the mismatch between the intensities of corresponding pixels in the two frames. (u, v) are the components of the motion vector to be determined for each pixel (x, y). $\mathrm{I}_0(x, y)$ is the intensity of the pixel at position (x, y) in the Reference frame. $\mathrm{I}_t(x + u, y + v)$ is the intensity of the pixel in the second frame, warped by the motion vector (u, v). $\mathcal{G}(\|x, y\|^2; \sigma) = e^{\frac{-\|x,y\|^2}{2\sigma^2}}$ is a Gaussian weighting function that gives higher importance to pixels closer to the center of the window. σ is the standard deviation of the Gaussian and the expression $\|x, y\|^2$ refers to the squared Euclidean norm of the spatial coordinates (x, y), which is equivalent to $x^2 + y^2$. This term is used to measure the distance of a pixel from the center of the window.

To derive the equations, we'll start by expressing the warped image intensity $\mathrm{I}_t(x + u, y + v)$ using a Taylor expansion around (x, y):

$$\mathrm{I}_t(x + u, y + v) \approx \mathrm{I}_0(x, y) + u \,.\frac{\partial \mathrm{I}_0}{\partial x} + v \,.\frac{\partial \mathrm{I}_0}{\partial y} \quad (3)$$

Now, substitute this expression into the energy function Eq. (2)

$$\mathcal{F}_t = \sum_{x,y} \left((\mathrm{I}_0(x, y) - [\mathrm{I}_0(x, y) + u \,.\frac{\partial \mathrm{I}_0}{\partial x} + v \,.\frac{\partial \mathrm{I}_0}{\partial y}] \right)^2 .\mathcal{G}(\left||x, y|\right|^2; \sigma) \quad (4)$$

Simplify and collect terms:

$$\mathcal{F}_t = \sum_{x,y} \left(u \,.\frac{\partial \mathrm{I}_0}{\partial x} + v \,.\frac{\partial \mathrm{I}_0}{\partial y} \right)^2 .\mathcal{G}\left(\left||x, y|\right|^2; \sigma \right) \quad (5)$$

Reference frame I_0 and optical flow $\mathcal{F}_t$. Output frame $\mathrm{I}_{Optical\_Flow\_image}$ = I_0. For each pixel (x, y) in I_0:

- Calculate the new position (x', y') using optical flow $\mathcal{F}_t : x' = x + v_{x,y}, y' = y + u_{x,y}$
- If $0 \leq x' < height(\mathrm{I}_0)\ and\ 0 \leq y' < width(\mathrm{I}_0)$:
  - Set $\mathrm{I}_{Optical\_Flow\_image}[x', y'] = \mathrm{I}_0[x, y]$.

$\mathrm{I}_{Optical\_Flow\_image}$ – The reference frame I_0 with optical flow-based transformations. Save the $\mathrm{I}_{Optical\_Flow\_image}$ –image into specified Emotion folder.

The Average contextual loss is calculated using feature frames, L1 loss, and optical flow. For each pixel in the reference frame feature I_{fr_0} and the corresponding pixel in the current frame feature I_{fr_t}, the L1 loss is computed. The loss is accumulated over all spatial positions, resulting in the Average contextual loss for each pair of feature frames.

The ACL for a current frame feature I_{fr_t} with respect to the reference frame feature $\mathrm{I}_{fr_0}$is given by:

$$L_{ACL}\left(\mathrm{I}_{fr\_0}, \mathrm{I}_{fr\_t}\right) = \frac{1}{N}\sum_{i,j}\left\|\mathcal{F}_t(x, y) \odot \left(\mathrm{I}_{fr\_0}(x, y) - \mathrm{I}_{fr\_t}(x + v, y + u)\right)\right\|_1 \quad (6)$$

$L_{ACL}\left(\mathrm{I}_{fr\_0}, \mathrm{I}_{fr\_t}\right)$: Average Contextual loss between the reference frame feature I_{fr_0} and the current frame feature I_{fr_t}. N: The total number of pixels in the image frames. It represents the normalization factor, ensuring that the loss is averaged over all pixels. $\sum_{i,j}$: Summation over all pixel positions in the frames. $\|.\|_1$: L1 norm, also known as the Manhattan norm or absolute norm. It is used to measure the absolute difference between corresponding pixel values.

$\mathcal{F}_t(x, y)$: Optical flow field at position (x, y) for the current frame feature I_{fr_t}. It represents the motion vector (displacement) of the pixel at position (x, y) between the reference frame feature and the current frame feature. $\odot$: Element-wise multiplication (Hadamard product) between the optical flow field $\mathcal{F}_t(x, y)$ and the absolute pixel-wise intensity difference $\mathrm{I}_{fr\_0}(x, y) - \mathrm{I}_{fr\_t}(x + u, y + v)$. This operation emphasizes regions where motion occurs. $\mathrm{I}_{fr\_0}(x, y)$: Intensity value of the pixel at position (x, y) in the reference frame feature. $\mathrm{I}_{fr\_t}(x + u, y + v)$: Intensity value of the pixel at the displaced position $(x + u, y + v)$ in the current frame feature. The displacement is determined by the optical flow vectors. Compute the frame contextual loss $L_{ACL}\left(\mathrm{I}_{fr\_0}, \mathrm{I}_{fr\_t}\right)$: using L1 loss and optical flow between reference frame features I_{fr_0} and current frame features I_{fr_t} after applying the temporally consistent flow smooth-flows. Append the Average contextual loss value and frame number to the Average contextual-losses list. Sort the list of Average contextual losses in descending order of loss values. Select the best four key pairs of frames that have the highest ACL value efficiently, and these selected frames are used to obtain resulting optical flow images.

The FCALF model's output for chosen optical flow images results in a dataset with fewer training instances. The deep learning algorithm, which is new and developing, cannot be used due to insufficient data. In order to artificially enhance the dataset size using several modification techniques like rotation, shifts, and flips, among others, image augmentation is usually necessary. The purpose of dataset enhancement is to reduce the likelihood of erroneous predictions resulting from over-fitting or overly rigorous pattern learning [41]. Therefore, each pixel in the resultant optical flow image shown in Fig. 3 is transformed to the new position in order to enhance the both appearance and dynamic changes in expressions results. This process is explained as follows:

The rotation changes the resultant image pixel (p, q) into a new rotation. $\theta = 20^0$ as $p' = p \times \cos(20^0) - q \times \sin(20^0); q' = p \times \sin(20^0) + q \times \cos(20^0)$ to obtain (p', q') the newly transformed coordinate. The next operation is translation, The width Trp and height Trq are moved to 20% in the following transformation to get the new expansion, which is expressed as $Tr'_p = p + Tr_p$ and height as $Tr'_q = q + Tr_q$, where Trp = 0.2 and Trq = 0.2. At the end, the horizontal Fhp and vertical Fvq flipping are performed on resultant image to get new transformation $Fh_p\ : p' = -p\ and\ q' = q$ and $Fv_q\ : p' = p\ and\ q' = -q$.

B. Transfer Learning Models

Training a large dataset with the current deep learning technique takes a week. To avoid time consumption, all researchers have used pre-trained models [42, 43]. Both overall error and training time are decreased by these pre-trained models. A portion of the models that have already been trained are used by fine-tuning the top layers in order to use them for the proposed approach. In addition, the pre-trained model's weights are frozen. The proposed approach uses two existing models, InceptionResNetV2[40] and VGG16 [33], and customizes the top layer by adding or removing layers and adjusting weights. The Keras API provides access to these models, which are used to identify emotions using augmented resultant optical flow image.

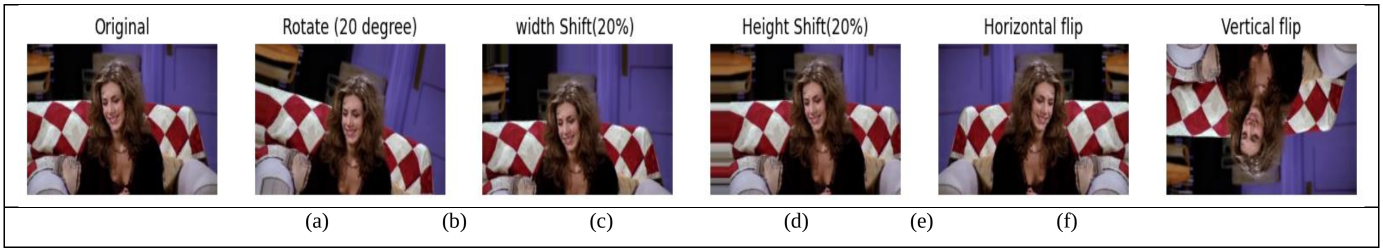

Fig. 3. Data augmentation operations on the resultant optical flow image of a sample happy emotion

C. Concatenation

Let's consider the pre-trained InceptionResNetV2 and VGG16 models are concatenated to create the ensemble models: INCRESV2-VGG16 as M, which is depicted pictorially in Fig. 2. In the experimental part, the effectiveness of hybrid models that have been pre-trained is covered in detail. The final dense layer of the InceptionResNetV2 and VGG16 model undergo further processing to generate ReLU activation functions, which are expressed as follows:

$$fpxn(x_i) = max(0, x_i)$$

Where, x_i represented by feature vector of different individual model InceptionResNetV2 (x_1) and VGG16 (x_2) are concatenated and those concatenated outputs are represented as follows:

Concatenation Function: $C([M]) = [fpxn(x_1), fpxn(x_2)]$

Where, $fxn(x_1)\ and\ fxn(x_2)$ are the feature vector output of the individual pre-trained model. The concatenated outputs of the base models are represented by M; to create an outcome vector of size k, where k represents the number of classes, the combined features vectors pass across a fully connected layer using a weight matrix W along with bias vector b. The completely linked layer's output is represented by the following.

$$z = W * [fpxn(x_1), fpxn(x_2)] + b$$

Next, the output vector z is subjected to the Softmax function in order to generate a probability distribution across all possible classes. The description of the Softmax function is:

$$P(y_i = 1|x) = \frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_j}}$$

where z is the i^{th} element of the outcome vector z and y_i is an indicator variable corresponding to the i^{th} class ($y_i = 1$ if the given input corresponds to class i^{th} and $y_i = 0$ otherwise). The total exponential of each element in the output vector is the denominator. Finally, for ensemble model, the Softmax function correctly identifies the emotion.

IV. Experiments, Results and Discussion

A. Datasets

For our experiments, we make use of the CAER-S dataset [10]. The dataset is well-suited for the task of emotion detection and focuses on context-aware emotion recognition. The dataset is comprised of video clips taken from 79 different television episodes. Each frame in the dataset is assigned to one of 7 emotional states: angry, disgusted, fearful, happy, sad, surprised, or neutral.

The Extended-Cohn-Kanade (CK+) dataset [44, 45] is a well-liked laboratory-controlled dataset of facial emotion detection. It consists of 327 images with labelling for 7 distinct emotion classes, taken from 118 distinct subjects of which 309 sequences have been labelled with six fundamental expressions using the FACS. The length of a video sequence, which can range from 10 to 60 images per second. Every video sequence starts with a neutral expression and ends with its most expressive face. Every video might have between 12 and 56 frames.

B. Flow Context Aware Loss Fusion (FCALF)

This FCALF model implements video frame processing using a VGG16 model for feature extraction and optical flow computation to assess Average contextual loss between frames, are discussed. Using PyTorch API, FCALF experiments are conducted on Google Colab GPUs. The description about the same is discussed as follows: the VGG16 model is loaded, focusing on the first 23 layers for feature extraction. Image preprocessing involves resizing to (224, 224) pixels and converting to PyTorch tensors. The FCALF model operates on an image sequence, using the first frame as the reference frame. Optical flow is calculated using Farneback's method, with parameters such as pyramid scale factor = 0.5, pyramid levels and no. of iterations at each pyramid level = 3, size of the pixel neighborhood used for Gaussian smoothing of the derivatives = 5, standard deviation of the Gaussian used for smoothing the derivatives = 1.2, and pixel neighborhood sizes used for Gaussian smoothing = 15 influencing the algorithm's sensitivity to motion and computational efficiency, and transformations are applied to the reference frame based on the flow. Average Contextual loss (ACL) is computed through L1 loss, frame features, and optical flow, with results printed for each frame, indicating the dissimilarity between features of the reference and current frames. The FCALF model selects best 4 key pairs of frames based on the highest ACL value for enhanced spatial-temporal analysis. The selected key pairs of frames are then used to obtain optical flow images. This high ACL value suggests a culmination of significant changes or the resolution of an emotional expression. Overall, the analysis underscores the dynamic nature of the video, with peaks in Average contextual loss values serving as markers for researchers to explore optical flow images of particular interest. The subtle understanding derived from these Average contextual losses enriches the scientific exploration of emotional dynamics in facial expressions within the video sequence, providing a quantitative basis for identifying and investigating key moments in the evolving emotional narrative.

The FCALF model apply on CAER-S datasets to extract best four key pairs of frames based on the highest ACL value for enhanced spatial-temporal analysis. The selected key pairs of frames are then used to obtain optical flow images, the provided data in Fig. 4 details the ACL values for each frame in a video sequence of Anger emotion, offering insights into the dynamic evolution of emotion. In the initial frames (1-5), ACL are relatively low, ranging from 18.8574 to 29.624, suggesting a period of stability or similarity with the reference frame. However, starting from Frame 6, there is a gradual increase in ACL, reaching 49.662 by Frame 10. This signifies a progression of dissimilarity, indicating potential shifts in emotional expression or notable changes in facial features. a notable spike occurs between Frames 13 and 14, where the ACL jumps from 58.9676 to 81.7295. This significant increase suggests a pivotal moment in the video, potentially capturing an intense emotional expression or distinct facial transformations. Subsequently, Frames 15 to 38 exhibit fluctuating ACL, reflecting a dynamic sequence with varying degrees of dissimilarity. In Fig. 5(a) Frames 23, 25, 26, and 28 shows the highest ACL values respectively, 121.5226, 122.6208, 120.6192, 124.7423, indicating instances of particularly pronounced differences. These resultant optical flow images are crucial for more detailed analysis, as they likely capture critical moments in the video sequence.

Similarly, for other Emotions in CAER-S shows in Fig. 4, For disgust (see Fig. 5(b)), the best four resultant optical flow images with the highest ACL values are frame numbers 45, 49, 61, and 48. The ACL values for these frames are 209.1001, 209.1548, 209.9282, and 211.1884, respectively. For fear (see Fig. 5(c)), the best four resultant optical flow images with the highest ACL values are frame numbers 11, 10, 25, and 9. The ACL values for these frames are 148.42, 148.6847, 149.2188, and 159.5878, respectively. for happy (see Fig. 5(d)), the best four resultant optical flow images with the highest ACL values are frame numbers 26, 27, 29, and 28. The ACL values for these frames are 135.8524, 136.5378, 139.2938, and 141.5067, respectively. for Neutral (see Fig. 5(e)), the best four resultant optical flow images with the highest ACL values are frame numbers 27, 33, 26, and 25. The ACL values for these frames are 167.9817, 168.0364, 173.2646, and 176.0626, respectively. For sadness (see Fig. 5(f)), the best four resultant optical flow images with the highest ACL values are frame numbers 6, 41, 7, and 40. The ACL values for these frames are 174.9732, 176.3151, 181.0617, and 183.9615, respectively. For surprise (see Fig. 5(g)), the best four resultant optical flow images with the highest ACL values are frame numbers 6, 18, 15, and 17. The ACL values for these frames are 93.2330, 95.5876, 95.8962, and 97.0862, respectively. Overall, the Fig. 4 and 5 show that the ACL values for all seven emotions in real time video sequence the highest ACL value shows the most significant changes of subtle emotion in video sequence. This suggests that these resultant optical flow images are the most informative for identifying the emotions.

The FCALF models also apply on CK+ dataset to extract the best 4 key pairs of frames based on the highest ACL value for enhanced spatial-temporal analysis. The selected key pairs of frames are then used to obtain optical flow images, The provided data in Fig. 6 details the ACL value for each frame in a video sequence of Anger emotion analysis reveals the dynamic evolution of a video sequence through distinct phases. Initially (Frame 1-5), frames show a gradual increase in average contextual loss (ACL), indicating a stable period with moderate dissimilarity from the reference frame. In the transition phase (Frame 6-10), there is a notable ACL increase, with Frame 10 standing out at 90.5806, suggesting a significant shift in facial features or emotional expression. Frames 11 to 14 depict a continuous rise in ACL, reaching 104.8176 in Frame 14, capturing sustained moments of emotional intensity or distinct facial transformations. Frame 15 marks a peak ACL value of 110.9438, followed by fluctuating values in Frames 16 to 20, suggesting dynamic changes and diverse emotional states, reflecting a dynamic sequence with varying degrees of dissimilarity. In Fig. 7(a) Frames 16, 18, 19, and 20 show the highest ACL values respectively, 115.0193, 116.5156, 117.1425, 117.9241, These resultant optical flow images likely capture the most intense moment of the anger expression.

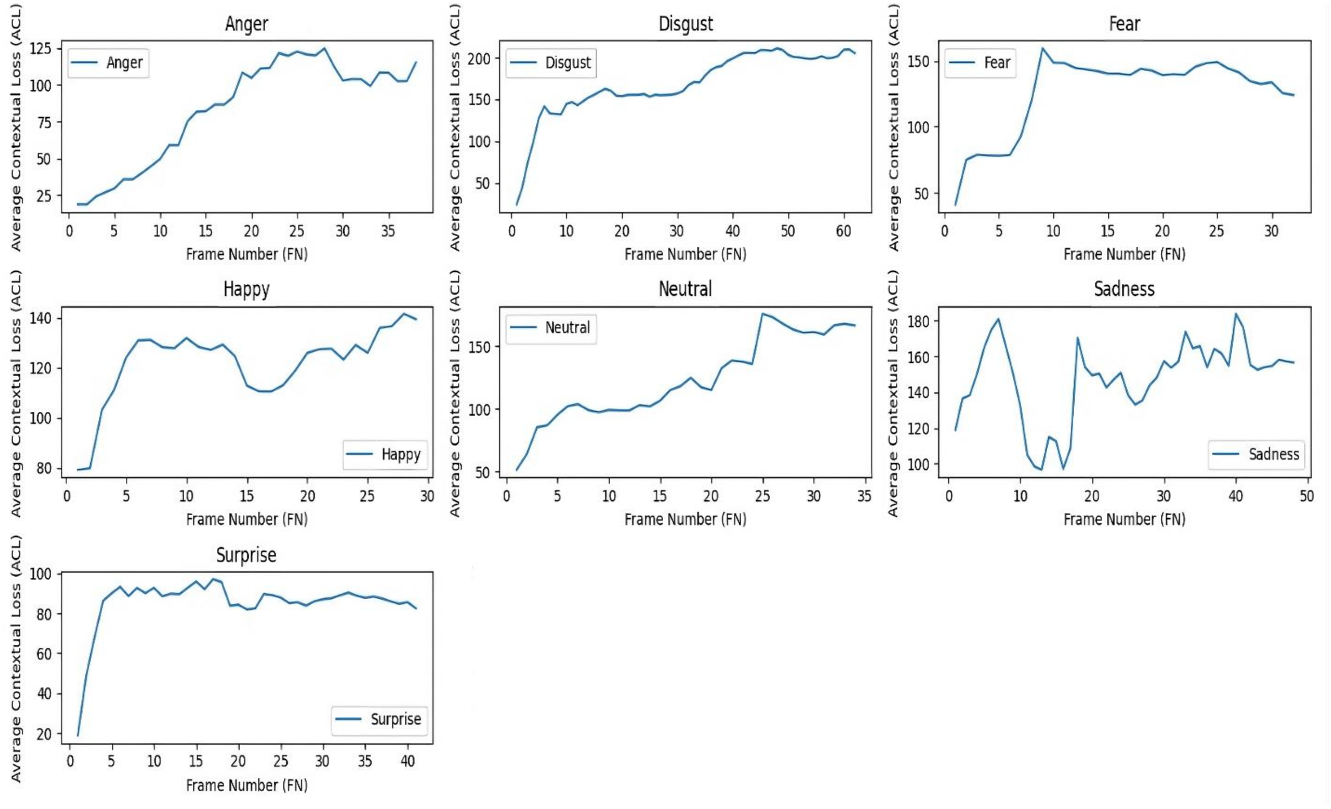

Fig. 4. For all seven emotions in the CAER-S dataset, the proposed FCALF model plots frame number (FN) with their average contextual loss (ACL) value.

Similarly, for other Emotions in CK+ shows in Fig. 6, for Disgust (see Fig. 7(b)): Frames 16, 17, 18, and 19 have the highest ACL values for these frames are 111.8592, 117.0162, 122.4170, and 124.8995. These resultant optical flow images might show the peak of the disgust emotion, with strong facial contortions. For Fear (see Fig. 7(c)): Frames 12, 11, 10, and 13 have the highest ACL values for these frames are 136.0939, 137.2431, 139.7179, and 142.8500. These resultant optical flow images likely depict the most frightened part of the fear emotion, with widened eyes and open mouths. For Happy (see Fig. 7(d)): Frames 9, 10, 11, and 12 have the highest ACL values for these frames are 93.5942, 97.9887, 101.7392, and 101.7757. These resultant optical flow images probably capture the broadest smiles and most outward emotion of joy. For Sadness (see Fig. 7(e)): Frames 21, 17, 19, and 18 have the highest ACL values for these frames are 117.2293, 117.9838, 119.3985, and 120.6697. These resultant optical flow images likely show the deepest sadness, with downcast eyes and furrowed brows. For Surprise (see Fig. 7(f)): Frames 10, 13, 12, and 11 have the highest ACL values for these frames are 111.7967, 112.4761, 113.3418, and 115.7291. These resultant optical flow images probably capture the moment of surprise, with raised eyebrows and open mouths.

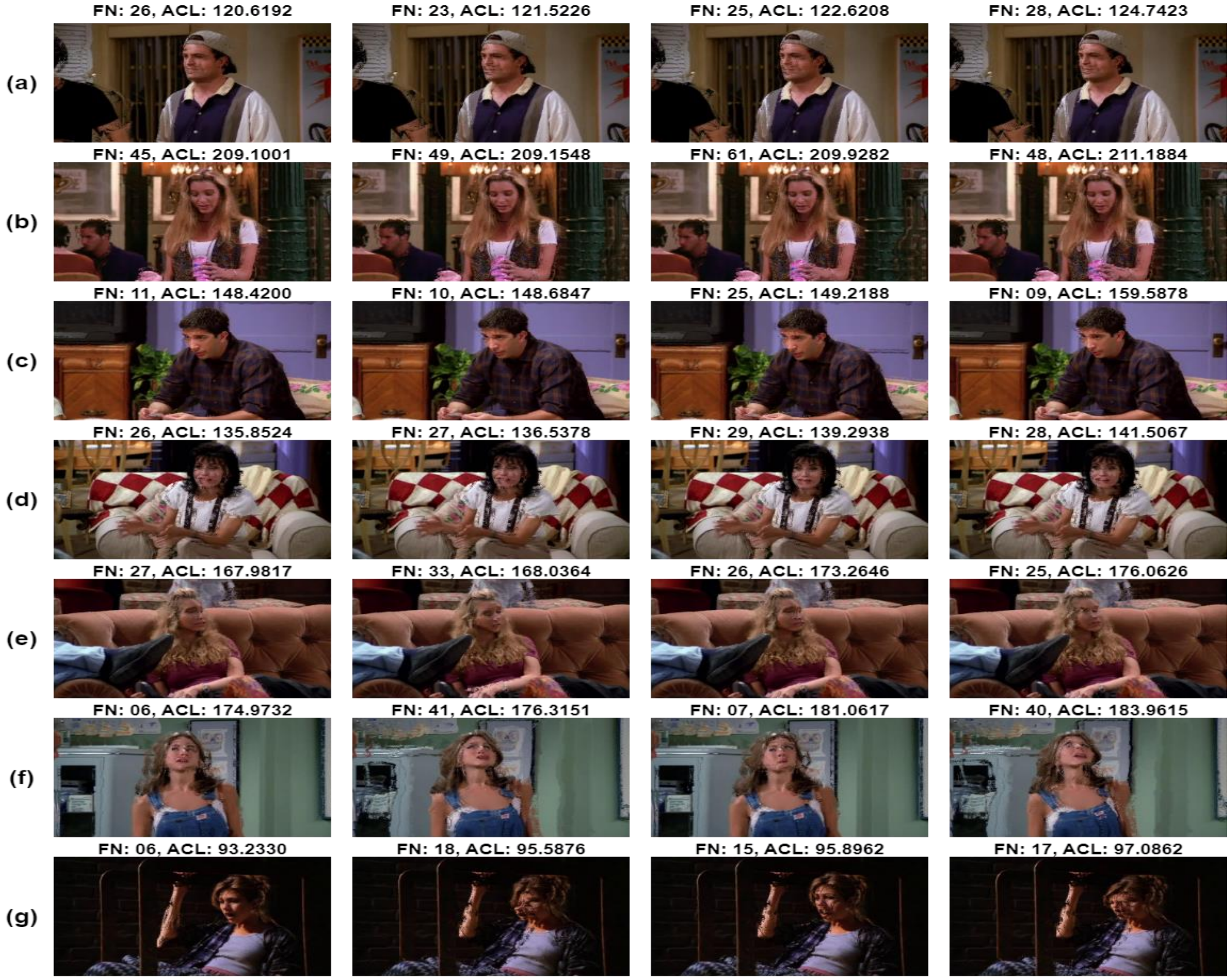

Fig. 5. The following are the sample resultant optical flow images of the selected best 4 key pairs of frames with Frame Number (FN) and highest Average Contextual Loss (ACL) values using the FCALF model on the CAER-S dataset: a) anger; b) disgust; c) fear; d) happiness; e) neutral; f) sadness; g) surprise

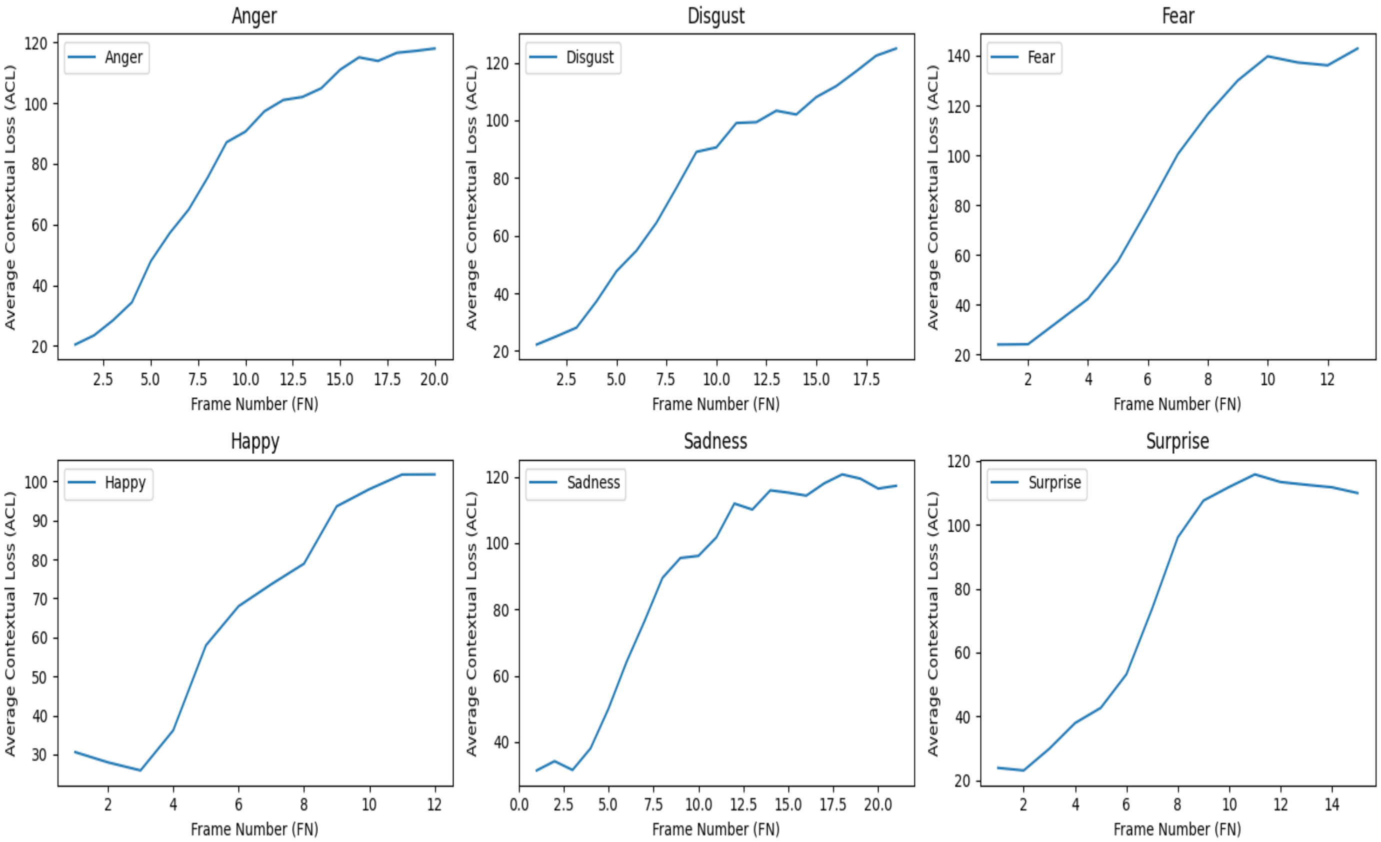

Fig. 6. For all six emotions in the CK+ dataset, the proposed FCALF model plots frame number (FN) with their average contextual loss (ACL) value

C. InceptionResNetV2 and VGG16

The FCALF model generated optical flow images are split into 50% training, 25% validation, and 25% testing sets. However, because the total number of images in training set is small after splitting, data-augmentation methods such as rotation (20^0), both width and height shift (0.2), and vertical as well as horizontal flipping are used to produce extra image variants. This would expand the amount of data in the training datasets and expose the pre-trained algorithms to additional image variants. TensorFlow Keras API is used to train the model, and experiments are conducted using Google colab GPUs. Pretrained models are InceptionResNetV2 and VGG16. The output images of FCALF model are fed into a pre-trained model, which is then fine-tuned to extract pertinent information from the output images of FCALF model. The section of content that follows discusses the pre-trained model's configuration setting.

The model, denoted as "model_2," comprises an input layer named "input_8" with a shape of (None, 224, 224, 3), indicating it takes images of size 224x224 pixels with three color channels (RGB). Two lambda layers, "lambda_12" and "lambda_15," transform the input data. The architecture integrates two pre-trained models: Inception-ResNet-v2, with 5x5 filters and 1,536 units in its last layer, and VGG16, with 7x7 filters and 512 units. Global max pooling layers, "global_max_pooling2d_11"and "global_max_pooling2d_12," follow each pre-trained model. Two separate dense layers, "dense_13" and "dense_16," with 128 units each, process the global max pooling outputs.

D. Concatenation

The resulting feature vectors are concatenated using a concatenate layer named "concatenate_1.", with 256 filters. Then the output of Concatenation layers is fed into Softmax layer to classified emotions. overall, this model integrates characteristics from Inception-ResNet-v2 and VGG16 to improve their representations for the purpose of classifying emotions into different emotion classes. The performance metrics of the proposed model are presented below.

E. Performance Measures

The performance matrix provided in Table II evaluates an emotion recognition model on the CAER-S dataset, presenting precision, recall, and F1-score metrics for individual emotion classes. Precision values, such as 0.96 for "Anger", "Happy", and "Sadness", signify the model's accuracy in predicting positive instances, with 96% of its predictions being accurate in the "Anger" class. Perfect recall scores in "Neutral".

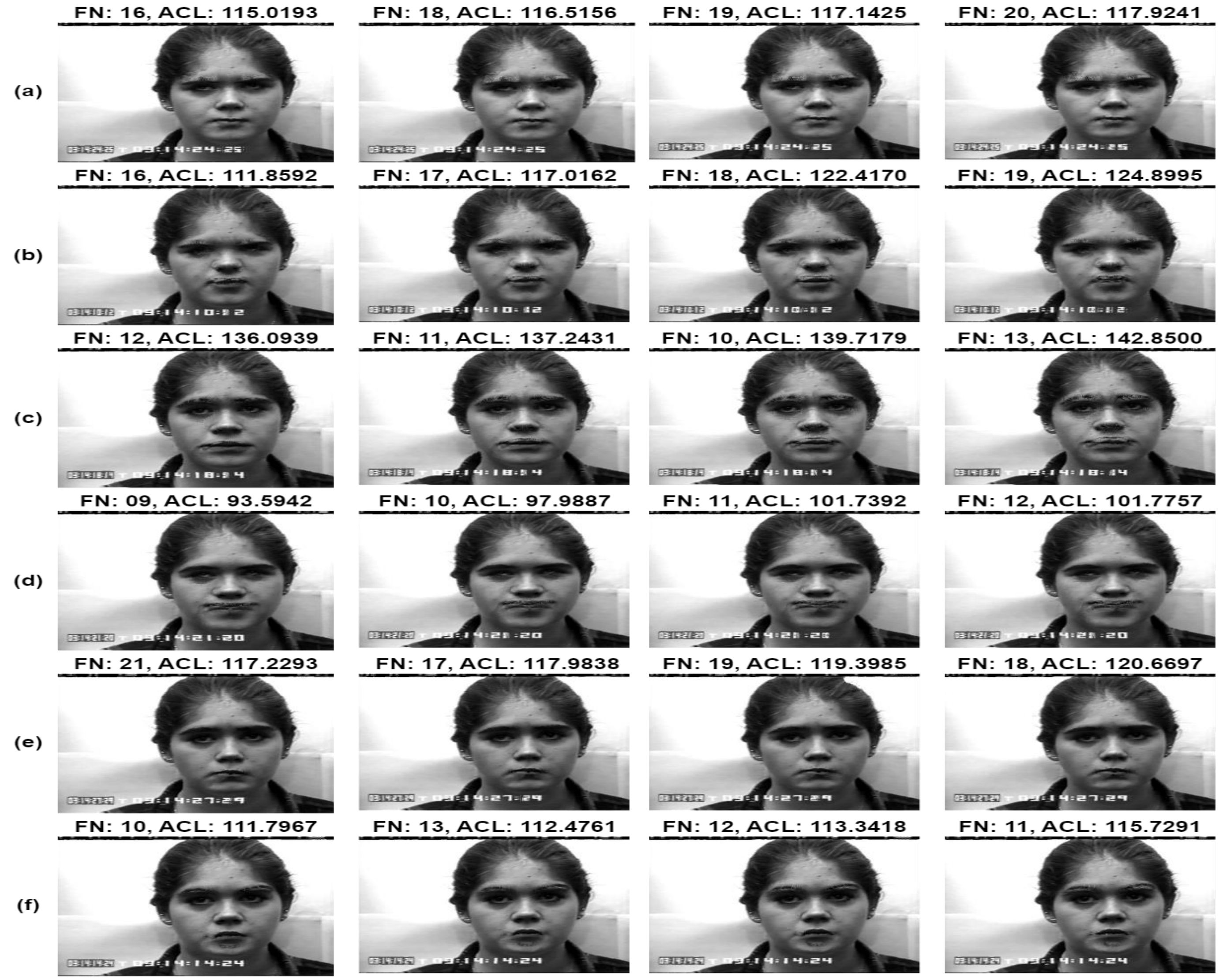

Fig. 7. The following are the sample resultant optical flow images of the selected best 4 key pairs of frames with Frame Number (FN) and highest Average Contextual Loss (ACL) values using the FCALF model on the CK+ dataset: a) anger; b) disgust; c) fear; d) happiness; e) sadness; f) surprise

"Happy", "Surprise" and "Sadness" indicate the model's ability to capture all instances of true positives within these classes, reaching 1.00 for the "Happy" class, signifying complete identification of Happy instances. F1-scores, such as 0.98 in "Disgust", "Fear", "Happy", and "Sadness," underscore the model's robust overall performance by balancing the trade-off between false positives and false negatives. In summary, the model demonstrates strong performance on the CAER-S dataset, with consistently high precision, recall, and F1-scores across diverse emotion classes, exemplifying its effectiveness in recognizing emotions within this dataset.

Similarly, the provided performance matrix in Table I offers a comprehensive assessment of an emotion recognition models proves on the CK+ dataset. This evaluation encompasses precision, recall, and F1-score metrics, providing a nuanced understanding of the model's predictive capabilities for distinct emotional classes. Precision, denoting the accuracy of positive predictions, is exemplified by the "Anger" class, where the model is correct 86% of the time. Recall, representing the model's ability to identify true positive instances, achieves perfection in the "Fear" class, indicating an adept capture of all instances of fear in the dataset. The F1-score, a harmonic mean of precision and recall, harmoniously balances these metrics and attains notable levels across classes. Noteworthy performances include flawless recognition in the "Fear" class and strong outcomes in "Surprise" with perfect precision and high recall, yielding an impressive F1-score of 0.98. Overall, the model exhibits commendable performance, demonstrating high precision, recall, and F1-scores across diverse emotional categories, affirming its effectiveness in recognizing facial expressions within the CK+ dataset.

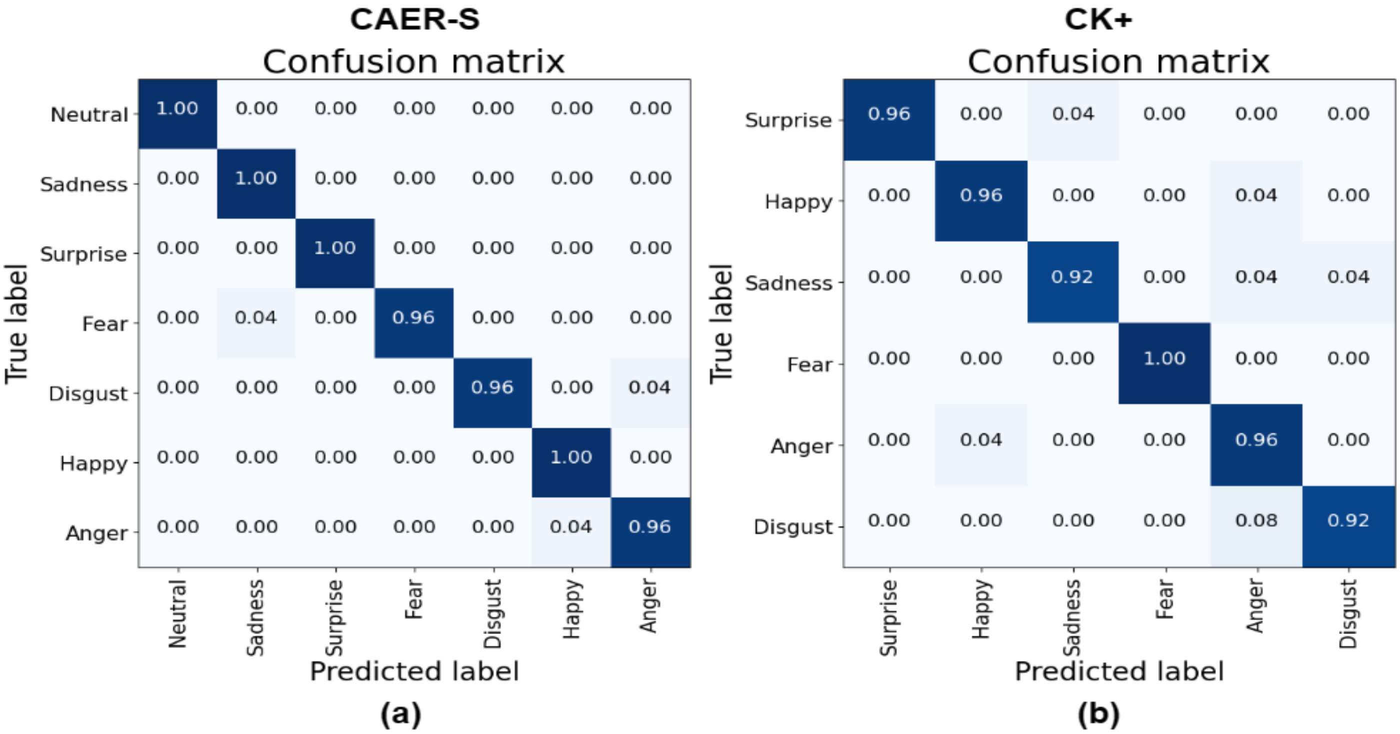

Fig. 8. The ERFN model's confusion matrix on the a) CAER-S and b) CK+ datasets

TABLE I. THE ERFN MODEL FOCUSES ON THE CLASSIFICATION PERFORMANCE PROPERTIES OF THE CK+ DATASET

Class	Precision	Recall	F1-Score
Anger	0.86	0.96	0.91
Disgust	0.96	0.92	0.94
Fear	1.00	1.00	1.00
Happy	0.96	0.96	0.96
Sadness	0.96	0.92	0.94
Surprise	1.00	0.96	0.98

TABLE II. THE ERFN MODEL FOCUSES ON THE CLASSIFICATION PERFORMANCE PROPERTIES OF THE CAER-S DATASET

Class	Precision	Recall	F1-Score
Anger	0.96	0.96	0.96
Disgust	1.00	0.96	0.98
Fear	1.00	0.96	0.98
Happy	0.96	1.00	0.98
Neutral	1.00	1.00	1.00
Sadness	0.96	1.00	0.98
Surprise	1.00	1.00	1.00

F. *Confusion Matrix*

The provided confusion matrix presents in Fig. 8(a) an evaluation of a proposed model on the CAER-S dataset, focusing on the recognition of seven emotions: Fear, Happy, Surprise, Sadness, Anger, Neutral, and Disgust. The diagonal elements indicate instances correctly classified for each emotion, revealing perfect accuracy for Surprise, Sadness, Neutral, and Happy. Disgust is recognized with high accuracy (96%), with a minor 4% misclassification into the Anger category. Similarly, Fear is identified with 96% accuracy, with a minor 4% misclassification into the Sadness category. Overall, the confusion matrix underscores the model's robust performance, particularly in distinguishing Neutral, Sadness, Happy, and Surprise emotions, but suggests a minor area for improvement in correctly identifying Fear, Anger, and Disgust emotions.

Similarly, the confusion matrix provided in Fig. 8(b) offers an evaluation of a proposed model's performance on the CK+ dataset, focusing on six facial expressions: Surprise, Happy, Sadness, Fear, Anger, and Disgust. Each row represents the true class, while each column corresponds to the predicted class. The diagonal elements of the matrix represent instances correctly classified for each expression, revealing high accuracy for Fear (100%), Surprise (96%), Happy (96%), and Sadness (92%). However, there are notable misclassifications, particularly between Anger and Disgust, with 8% of Disgust instances mistakenly predicted as Anger. Additionally, 4% of Sadness expressions are misclassified as both Anger and Disgust. Overall, while the model demonstrates commendable accuracy for certain expressions, the confusion matrix highlights areas for improvement, particularly in distinguishing between Anger and Disgust, and refining the model's recognition of Sadness expressions.

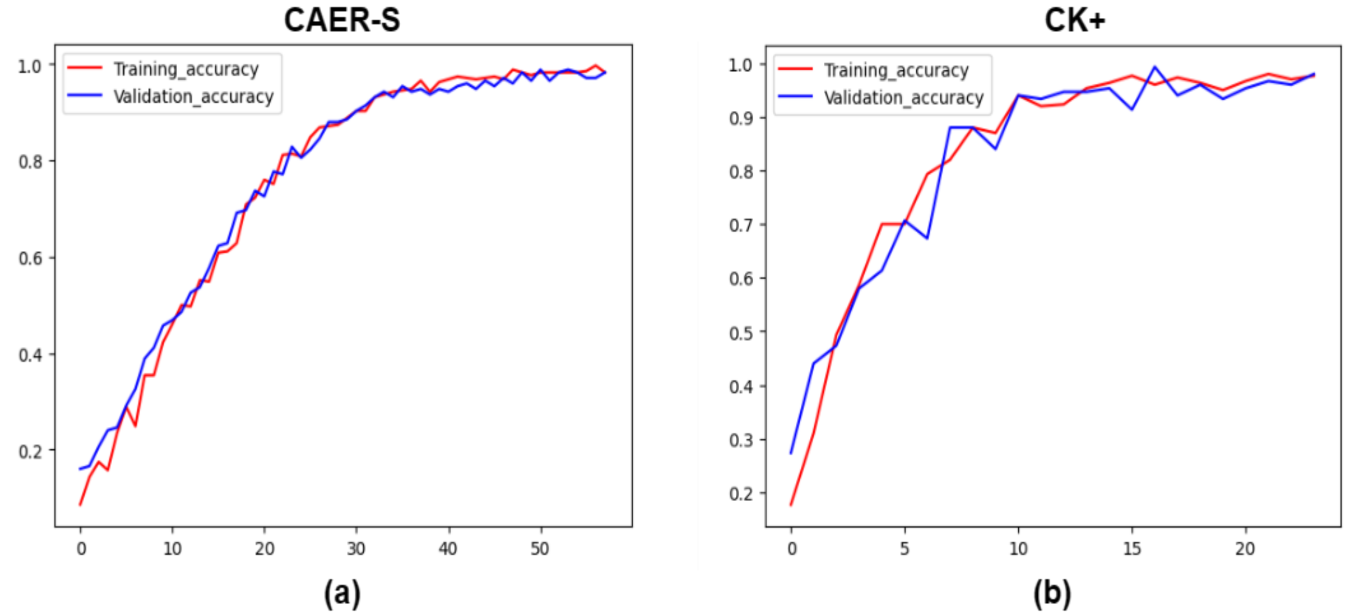

Fig. 9. The ERFN model's training and validation accuracy was assessed using the a) CAER-S and b) CK+ datasets

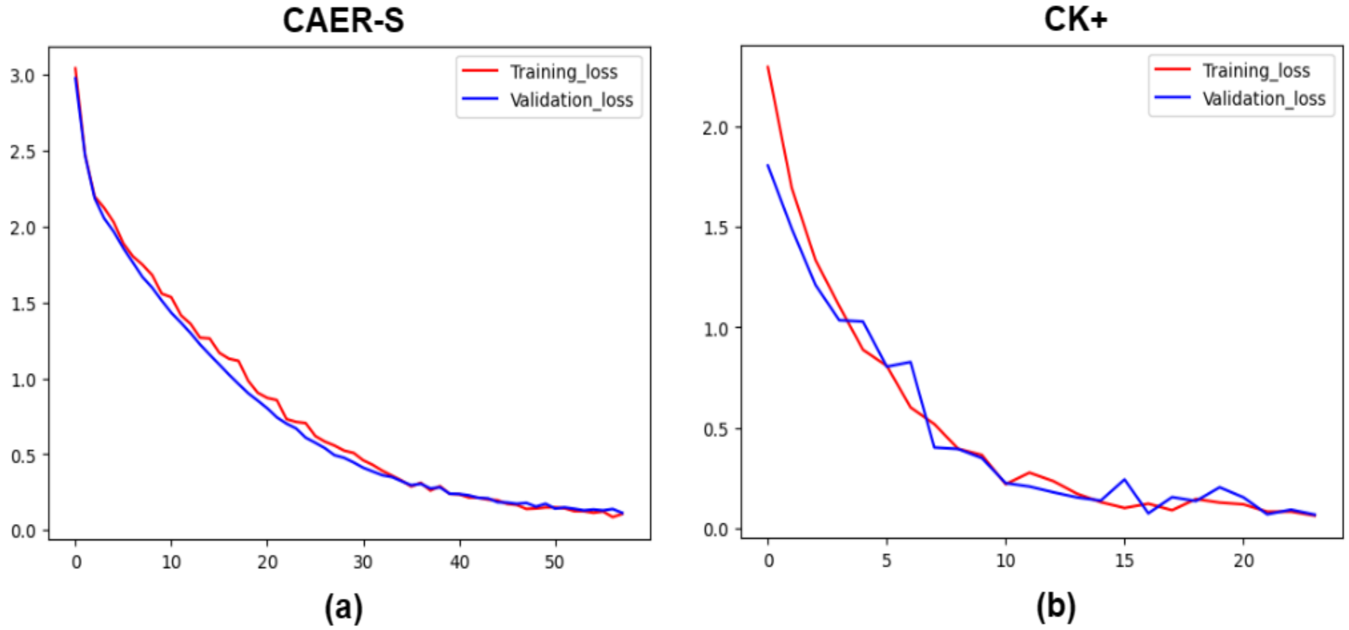

Fig. 10. The ERFN model's training and validation loss was assessed using the a) CAER-S and b) CK+ datasets

G. Comparison to State-of-the-art Methods

The ERFN model outperforms state-of-the-art techniques on the CAER-S and CK+ datasets, demonstrating a substantial improvement in emotion identification. The ERFN model obtains a superior accuracy of 98.29% on the CAER-S dataset shown in Table III, outperforming previous models with accuracies ranging from 73.51% to 93.26%. The ERFN model obtains a 98.00% accuracy on the CK+ dataset shown in Table III, surpassing previous techniques that have accuracies ranging from 87.16% to 97.79%. The results highlight the effectiveness and strength of our proposed method, showcasing its superiority in reliably identifying emotions across various datasets.

H. Visualization using Grad-Cam

The qualitative outcomes of trained Grad-Cam maps produced by Grad-CAM [75] using optimized VGG16Net are displayed in Fig. 11. It should be noted that images in Fig. 11 were accurately identified using refined VGG16Net to ground truth emotion categories. In the CAER-S dataset, Grad-Cam effectively localizes context information, which can improve the performance of emotion identification in a context-aware model.

TABLE III. THE STUDY FOCUSES ON THE TWO PROVIDED DATASETS AND COMPARES THEM WITH ALTERNATIVE METHODOLOGIES

Datasets	Methodology	Accuracy (%)
CAER-S	CAER-Net-S[10] CAAGR[28] MobileNet-V2[46] MHCAN[47] Attention-Guided-CAESR[48] GCN-In-Context[49] ResNet-18[39] Body-Object Attention (BDA)[32] Res2Net-50[50] EfficientFace[51] MATF[52] MA-Net[53] GLAMOR-Net[54] CAER-VRD[55] HCBER-with-Scene-Graph[56] Hierarchical Attention Module (HAM)[57] GFFT[58] CD-Net[59] ERFN(Our)	73.51 77.02 79.23 79.64 81.00 81.31 84.67 84.82 85.35 85.87 86.11 88.42 89.88 90.49 90.83 92.86 92.98 93.26 98.29
CK+	IT-RBM[60] GM-WLBP+GLCMRM+CNN-LSTM[61] LGC-HD[62] Optical Flow Reconstruction[63] LPQ-LBP- HOG- MSVM[64] DCNN-HSA[65] SCD Learning[66] ExNet[67] DAM-CNN[68] EIFN[69] RASnet-ERSnet-MAblocks[70] Multi-modal + EEG + BiLSTM[71] Improved-RNN[72] RGCFace[73] SISTCM-TLSTM[74] ERFN(Our)	87.16 91.42 92.30 92.80 94.20 95.71 95.73 95.81 95.88 96.02 96.28 96.36 96.37 97.30 97.79 98.00

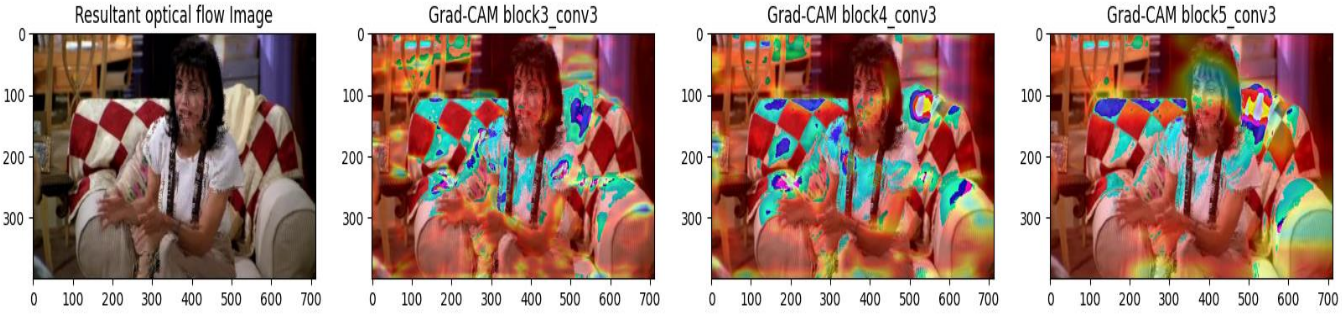

Fig. 11. Sample resultant optical flow image of a happy emotion in CAER-S dataset along with Grad-CAM maps in different layers

I. Training and Validation Graph

The Proposed model is trained separately using training, validation and test set on CAER-S and CK+ Datasets. In the proposed model on CAER-S and CK+ Dataset, Categorical-cross entropy loss function and the Nadam optimizer are used with a learning rate of 0.00001 and 0.0001 respectively. The Proposed model is trained with batch size of 32 for 80 epochs. To avoid overfitting, this epoch value (80) will be stopped earlier. In the experiment the proposed model used early stopping according to the training and validation accuracy of the proposed model on CAER-S and CK+ Dataset stopped improving after 58 and 24 epochs respectively which is shown in Fig. 9 (a and b). It is observed from the Fig. 9(a) that the validation (98.29%) and training (98.29%) accuracy both are same and also observed from the Fig. 9(b) that the validation (98%) and training (97.67%) accuracy. This means the proposed model is able to classified emotion for new data. When the validation and training accuracies reach the same value, the training process is stopped early to avoid overfitting. Early stopping allows the model to train for fewer epochs, which can save time and computational resources. Fig. 10(a) and (b) illustrates the observed loss performance results for the proposed model. The loss for each epoch is shown on the epoch vs. loss graph. As epochs increase, loss values decrease, as shown in Fig. 10(a) and (b). It is observed from the Fig. 10 (a) and (b) that the validation loss and training loss has very small gap and low loss. This means the proposed model is performing well on both the training and validation set, and is likely to generalize well to unseen data. It is observed from the Fig. 10(a) that the training (0.1067) and validation (0.1136) loss and also observed from the Fig. 10(b) that the training (0.0622) and validation (0.0670) loss. Out of all the state-of-the-art model, ERFN model performed the best in terms of Accuracy, loss, Precision, recall, and F1-score.

V. Conclusion and Future Work

The proposed work looks into the problems with current methods of emotions recognition that relies mostly on facial movements. The model suggests the Emotion Recognition Fusion Network (ERFN), a new model that uses body and contextual information to make up for the lack of specific facial cues. Advanced methods are used in the ERFN process. One of these is the Flow Context Aware Loss Fusion (FCALF) model. This model uses deep feature extraction (using VGG16), Farneback optical flow analysis, and L1 loss to find the Average Contextual Loss (ACL). Finding the four pairs of frames with the highest ACL values, getting optical flow images from these frames, and improving model generalization through pre-trained model are the most important parts of our method. We fine-tune both InceptionResNetV2 and VGG16 models, incorporating GlobalMaxPool2D and Dense layers to capture intricate details and flow-contextual information from face, body, and scene. We make strong feature representations by joining the results from these models together. According to the results of our experiments, the ERFN is more accurate and useful than other models. It is particularly effective at picking up on context-aware emotions, making it effective in real-world uncontrolled environments. The proposed method could help make emotions recognition technology better. Future work will focus on developing and integrating audio processing techniques to analyze speech and vocal tones, which, when combined with visual data, can significantly enhance the model's performance in real-world applications.

Acknowledgment

In terms of problem formulation, solutions, literature review, data corrections and interpretations of findings, all authors are contributed equally in the present work.

Data Availability

The Extended Cohen Kanade (CK+) dataset can be distributed free of charge for research purposes and non-commercial use only, and one can send the request to mer160@pitt.edu for a downloading link.

References

[1] S. Li and W. Deng, "Deep facial expression recognition: A survey," IEEE transactions on affective computing, vol. 13, no. 3, pp. 1195-1215, 2020.

[2] S. Lugović, I. Dunđer, and M. Horvat, "Techniques and applications of emotion recognition in speech," in 2016 39th international convention on information and communication technology, electronics and microelectronics (mipro), 2016, pp. 1278-1283: IEEE.

[3] D. Dangi, A. Bhagat, and D. K. Dixit, "Sentiment Analysis on Social Media Using Genetic Algorithm with CNN," Computers, Materials & Continua, vol. 70, no. 3, 2022.

[4] H. Yang, U. Ciftci, and L. Yin, "Facial expression recognition by de-expression residue learning," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 2168-2177.

[5] Z. Pan, Y. Wang, and S. Zhang, "Joint face detection and Facial Landmark Localization using graph match and pseudo label," Signal Processing: Image Communication, vol. 102, p. 116587, 2022.

[6] D. Liu, X. Ouyang, S. Xu, P. Zhou, K. He, and S. Wen, "SAANet: Siamese action-units attention network for improving dynamic facial expression recognition," Neurocomputing, vol. 413, pp. 145-157, 2020.

[7] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, "Context based emotion recognition using emotic dataset," IEEE transactions on pattern analysis and machine intelligence, vol. 42, no. 11, pp. 2755-2766, 2019.

[8] H. Aviezer, Y. Trope, and A. Todorov, "Body cues, not facial expressions, discriminate between intense positive and negative emotions," Science, vol. 338, no. 6111, pp. 1225-1229, 2012.

[9] J. K. McNulty and F. D. Fincham, "Beyond positive psychology? Toward a contextual view of psychological processes and well-being," American Psychologist, vol. 67, no. 2, p. 101, 2012.

[10] J. Lee, S. Kim, S. Kim, J. Park, and K. Sohn, "Context-aware emotion recognition networks," in Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 10143-10152.

[11] W. Zheng, "Multi-view facial expression recognition based on group sparse reduced-rank regression," IEEE Transactions on Affective Computing, vol. 5, no. 1, pp. 71-85, 2014.

[12] Z. Li, J.-i. Imai, and M. Kaneko, "Facial-component-based bag of words and phog descriptor for facial expression recognition," in 2009 IEEE International Conference on Systems, Man and Cybernetics, 2009, pp. 1353-1358: IEEE.

[13] A. Dapogny, K. Bailly, and S. Dubuisson, "Dynamic facial expression recognition by joint static and multi-time gap transition classification," in 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2015, vol. 1, pp. 1-6: IEEE.

[14] S. Li, W. Deng, and J. 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, 2017, pp. 2852-2861.

[15] 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 Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 5562-5570.

[16] L. Zhong, Q. Liu, P. Yang, J. Huang, and D. N. Metaxas, "Learning multiscale active facial patches for expression analysis," IEEE transactions on cybernetics, vol. 45, no. 8, pp. 1499-1510, 2014.

[17] I. Kotsia and I. Pitas, "Facial expression recognition in image sequences using geometric deformation features and support vector machines," IEEE transactions on image processing, vol. 16, no. 1, pp. 172-187, 2006.

[18] H.-D. Nguyen, S.-H. Kim, G.-S. Lee, H.-J. Yang, I.-S. Na, and S.-H. Kim, "Facial expression recognition using a temporal ensemble of multi-level convolutional neural networks," IEEE Transactions on Affective Computing, vol. 13, no. 1, pp. 226-237, 2019.

[19] K. Karpouzis et al., "Modeling naturalistic affective states via facial, vocal, and bodily expressions recognition," in Artifical Intelligence for Human Computing: ICMI 2006 and IJCAI 2007 International Workshops, Banff, Canada, November 3, 2006, Hyderabad, India, January 6, 2007, Revised Seleced and Invited Papers, 2007, pp. 91-112: Springer.

[20] M. A. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," IEEE Transactions on Affective Computing, vol. 2, no. 2, pp. 92-105, 2011.

[21] K. Schindler, L. Van Gool, and B. De Gelder, "Recognizing emotions expressed by body pose: A biologically inspired neural model," Neural networks, vol. 21, no. 9, pp. 1238-1246, 2008.

[22] Z. Yang and S. S. Narayanan, "Modeling dynamics of expressive body gestures in dyadic interactions," IEEE Transactions on Affective Computing, vol. 8, no. 3, pp. 369-381, 2016.

[23] P. Barros, D. Jirak, C. Weber, and S. Wermter, "Multimodal emotional state recognition using sequence-dependent deep hierarchical features," Neural Networks, vol. 72, pp. 140-151, 2015.

[24] P. Barros, G. I. Parisi, C. Weber, and S. Wermter, "Emotion-modulated attention improves expression recognition: A deep learning model," Neurocomputing, vol. 253, pp. 104-114, 2017.

[25] D. Nguyen, K. Nguyen, S. Sridharan, D. Dean, and C. Fookes, "Deep spatio-temporal feature fusion with compact bilinear pooling for

multimodal emotion recognition," Computer Vision and Image Understanding, vol. 174, pp. 33-42, 2018.

[26] L. F. Barrett, B. Mesquita, and M. Gendron, "Context in emotion perception," Current directions in psychological science, vol. 20, no. 5, pp. 286-290, 2011.

[27] Z. Chen and D. Whitney, "Tracking the affective state of unseen persons," Proceedings of the National Academy of Sciences, vol. 116, no. 15, pp. 7559-7564, 2019.

[28] M. Zhang, Y. Liang, and H. Ma, "Context-aware affective graph reasoning for emotion recognition," in 2019 IEEE International Conference on Multimedia and Expo (ICME), 2019, pp. 151-156: IEEE.

[29] T. Mittal, P. Guhan, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "Emoticon: Context-aware multimodal emotion recognition using frege's principle," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 14234-14243.

[30] A. Bandyopadhyay, S. Sarkar, A. Mukherjee, S. Bhattacherjee, and S. Basu, "Identifying emotional facial expressions in practice: A study on medical students," Indian Journal of Psychological Medicine, vol. 43, no. 1, pp. 51-57, 2021.

[31] S. D'mello and A. Graesser, "AutoTutor and affective AutoTutor: Learning by talking with cognitively and emotionally intelligent computers that talk back," ACM Transactions on Interactive Intelligent Systems (TiiS), vol. 2, no. 4, pp. 1-39, 2013.

[32] W. Li, X. Dong, and Y. Wang, "Human emotion recognition with relational region-level analysis," IEEE Transactions on Affective Computing, vol. 14, no. 1, pp. 650-663, 2021.

[33] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[34] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in Image Analysis: 13th Scandinavian Conference, SCIA 2003 Halmstad, Sweden, June 29–July 2, 2003 Proceedings 13, 2003, pp. 363-370: Springer.

[35] X. Song, Y. Zhao, and J. Yang, "STC-Flow: Spatio-temporal context-aware optical flow estimation," Signal Processing: Image Communication, vol. 99, p. 116441, 2021.

[36] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," Advances in neural information processing systems, vol. 27, 2014.

[37] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 4489-4497.

[38] S. Zhou, X. Wu, F. Jiang, Q. Huang, and C. Huang, "Emotion recognition from large-scale video clips with cross-attention and hybrid feature weighting neural networks," International Journal of Environmental Research and Public Health, vol. 20, no. 2, p. 1400, 2023.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770-778.

[40] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in Proceedings of the AAAI conference on artificial intelligence, 2017, vol. 31, no. 1.

[41] A. Mumuni and F. Mumuni, "Data augmentation: A comprehensive survey of modern approaches," Array, p. 100258, 2022.

[42] M. Iman, H. R. Arabnia, and K. Rasheed, "A review of deep transfer learning and recent advancements," Technologies, vol. 11, no. 2, p. 40, 2023.

[43] N. D. Kathamuthu et al., "A deep transfer learning-based convolution neural network model for COVID-19 detection using computed tomography scan images for medical applications," Advances in Engineering Software, vol. 175, p. 103317, 2023.

[44] 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 computer society conference on computer vision and pattern recognition-workshops, 2010, pp. 94-101: IEEE.

[45] T. Kanade, J. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in Proceedings fourth IEEE international conference on automatic face and gesture recognition (cat. No. PR00580), 2000, pp. 46-53: IEEE.

[46] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 4510-4520.

[47] Y. Yuan, F. Lu, X. Cheng, and Y. Liu, "Context Based Vision Emotion Recognition in the Wild," in 2022 IEEE 17th Conference on Industrial Electronics and Applications (ICIEA), 2022, pp. 479-484: IEEE.

[48] S. Jaiswal, S. Misra, and G. Nandi, "Attention-guided context-aware emotional state recognition," in 2020 IEEE 7th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON), 2020, pp. 1-6: IEEE.

[49] H. Zeng, G. Li, T. Tong, and Q. Gao, "A graph convolutional network for emotion recognition in context," in 2020 Cross Strait Radio Science & Wireless Technology Conference (CSRSWTC), 2020, pp. 1-3: IEEE.

[50] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2net: A new multi-scale backbone architecture," IEEE transactions on pattern analysis and machine intelligence, vol. 43, no. 2, pp. 652-662, 2019.

[51] Z. Zhao, Q. Liu, and F. Zhou, "Robust lightweight facial expression recognition network with label distribution training," in Proceedings of the AAAI conference on artificial intelligence, 2021, vol. 35, no. 4, pp. 3510-3519.

[52] Y. Guo et al., "Facial expressions recognition with multi-region divided attention networks for smart education cloud applications," Neurocomputing, vol. 493, pp. 119-128, 2022.

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

[54] N. Le, K. Nguyen, A. Nguyen, and B. Le, "Global-local attention for emotion recognition," Neural Computing and Applications, vol. 34, no. 24, pp. 21625-21639, 2022.

[55] M.-H. Hoang, S.-H. Kim, H.-J. Yang, and G.-S. Lee, "Context-aware emotion recognition based on visual relationship detection," IEEE Access, vol. 9, pp. 90465-90474, 2021.

[56] S. Wu, L. Zhou, Z. Hu, and J. Liu, "Hierarchical context-based emotion recognition with scene graphs," IEEE Transactions on Neural Networks and Learning Systems, 2022.

[57] H. Tao and Q. Duan, "Hierarchical attention network with progressive feature fusion for facial expression recognition," Neural Networks, vol. 170, pp. 337-348, 2024.

[58] R. Xu, A. Huang, Y. Hu, and X. Feng, "GFFT: Global-local feature fusion transformers for facial expression recognition in the wild," Image and Vision Computing, vol. 139, p. 104824, 2023.

[59] Z. Wang, L. Lao, X. Zhang, Y. Li, T. Zhang, and Z. Cui, "Context-dependent emotion recognition," Journal of Visual Communication and Image Representation, vol. 89, p. 103679, 2022.

[60] S. Wang, Z. Zheng, S. Yin, J. Yang, and Q. Ji, "A novel dynamic model capturing spatial and temporal patterns for facial expression analysis," IEEE transactions on pattern analysis and machine intelligence, vol. 42, no. 9, pp. 2082-2095, 2019.

[61] Z. Ullah, L. Qi, D. Binu, B. Rajakumar, and B. Mohammed Ismail, "2-D canonical correlation analysis based image super-resolution scheme for facial emotion recognition," Multimedia Tools and Applications, vol. 81, no. 10, pp. 13911-13934, 2022.

[62] D. G. R. Kola and S. K. Samayamantula, "Facial expression recognition using singular values and wavelet-based LGC-HD operator," IET Biometrics, vol. 10, no. 2, pp. 207-218, 2021.

[63] D. Poux, B. Allaert, N. Ihaddadene, I. M. Bilasco, C. Djeraba, and M. Bennamoun, "Dynamic facial expression recognition under partial occlusion with optical flow reconstruction," IEEE Transactions on Image Processing, vol. 31, pp. 446-457, 2021.

[64] N. Kumar HN, A. S. Kumar, G. Prasad MS, and M. A. Shah, "Automatic facial expression recognition combining texture and shape features from

prominent facial regions," IET Image Processing, vol. 17, no. 4, pp. 1111-1125, 2023.

[65] C. Gan, J. Xiao, Z. Wang, Z. Zhang, and Q. Zhu, "Facial expression recognition using densely connected convolutional neural network and hierarchical spatial attention," Image and vision computing, vol. 117, p. 104342, 2022.

[66] A. B. Tanfous, H. Drira, and B. B. Amor, "Sparse coding of shape trajectories for facial expression and action recognition," IEEE transactions on pattern analysis and machine intelligence, vol. 42, no. 10, pp. 2594-2607, 2019.

[67] M. N. Riaz, Y. Shen, M. Sohail, and M. Guo, "Exnet: An efficient approach for emotion recognition in the wild," Sensors, vol. 20, no. 4, p. 1087, 2020.

[68] S. Xie, H. Hu, and Y. Wu, "Deep multi-path convolutional neural network joint with salient region attention for facial expression recognition," Pattern recognition, vol. 92, pp. 177-191, 2019.

[69] H. Zhang, W. Su, and Z. Wang, "Expression-identity fusion network for facial expression recognition," in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 2122-2126: IEEE.

[70] Y. Gan, J. Chen, Z. Yang, and L. Xu, "Multiple attention network for facial expression recognition," IEEE Access, vol. 8, pp. 7383-7393, 2020.

[71] Y. Wu and J. Li, "Multi-modal emotion identification fusing facial expression and EEG," Multimedia Tools and Applications, vol. 82, no. 7, pp. 10901-10919, 2023.

[72] W. Zhang, X. Zhang, and Y. Tang, "Facial expression recognition based on improved residual network," IET Image Processing, 2023.

[73] Q. Liu, Y. Zhou, W. Liu, G. Li, C. Li, and W. Chen, "RGCFace: Regularized Global Center loss for Deep Facial Expression Recognition," in 2022 4th International Conference on Data Intelligence and Security (ICDIS), 2022, pp. 292-297: IEEE.

[74] J. Wei, G. Hu, X. Yang, A. T. Luu, and Y. Dong, "Learning facial expression and body gesture visual information for video emotion recognition," Expert Systems with Applications, vol. 237, p. 121419, 2024.

[75] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2921-2929.