Facial Expression Recognition with an Efficient Mix Transformer for Affective Human-Robot Interaction

Farshad Safavi, *Student Member, IEEE*, Kulin Patel, *Student Member, IEEE*, and Ramana Vinjamuri, *Senior Member, IEEE*

***Abstract*—Emotion recognition can significantly enhance interactions between humans and robots, particularly in shared tasks and collaborative processes. Facial Expression Recognition (FER) allows affective robots to adapt their behavior in a socially appropriate manner. However, the potential of efficient Transformers for FER remains underexplored. Additionally, leveraging self-attention mechanisms to create segmentation masks that accentuate facial landmarks for improved accuracy has not been fully investigated. Furthermore, current FER methods lack computational efficiency and scalability, limiting their applicability in real-time scenarios. Therefore, we developed the robust, scalable, and generalizable EmoFormer model, incorporating an efficient Mix Transformer block along with a novel fusion block. Our approach scales across a range of models from EmoFormer-B0 to EmoFormer-B2. The main innovation lies in the fusion block, which uses element-wise multiplication of facial landmarks to emphasize their role in the feature map. This integration of local and global attention creates powerful representations. The efficient self-attention mechanism within the Mix Transformer establishes connections among various facial regions. It enhances efficiency while maintaining accuracy in emotion classification from facial landmarks. We evaluated our approach for both categorical and dimensional facial expression recognition on four datasets: FER2013, AffectNet-7, AffectNet-8, and DEAP. Our ensemble method achieved state-of-the-art results, with accuracies of 77.35% on FER2013, 67.71% on AffectNet-7, and 65.14% on AffectNet-8. For the DEAP dataset, our method achieved 98.07% accuracy for arousal and 97.86% for valence, demonstrating the robustness and generalizability of our models. As an application of our method, we implemented EmoFormer in an affective robotic arm, enabling the human-robot interaction system to adjust its speed based on the user's facial expressions. This was validated through a user experiment with six subjects, demonstrating the feasibility and effectiveness of our approach in creating emotionally intelligent human-robot interactions. Overall, our results demonstrate that EmoFormer is a robust, efficient, and scalable solution for FER, with significant potential for advancing human-robot interaction through emotion-aware robotics.**

***Index Terms*—Affective computing, deep learning, facial expression recognition, human-robot interaction, transformer.**

I. Introduction

EMOTIONS play a crucial role in human-robot interaction, enabling universal social communication and aiding decision-making in robotic machines. Nonverbal elements convey emotional messages in human communication. Therefore, understanding affective states such as facial expressions or emotions expressed via bodily gestures provides valuable information to an emotional robot about human intent. In this work, we exclusively developed a robust facial expression recognition model for human-robot interaction. The main objective of facial expression recognition is to identify the emotional state of a person by extracting relevant facial features from images and videos. FER classifies emotions into two types of representations. First, categorical representations divide emotions into distinct states such as happy, sad, angry, etc [1]. On the other hand, dimensional representations map emotions onto a multi-dimensional space, offering a continuous measure of arousal and valence. Valence, which is a pleasantness continuum, indicates whether the emotion is positive or negative, while arousal refers to the intensity of the emotion [2]. In this research, we test our model across both categorical and dimensional datasets.

Fig. 1. Row (A) shows the original images, and Row (B) shows the EmoFormer facial landmark identifications on the AffectNet Dataset. The blue dots represent facial landmarks detected by our new EmoFormer model. EmoFormer identifies key facial features, such as regions around the eyes and mouth, marking these areas with light blue while excluding less relevant regions like the hair and jawline.

The main challenges that FER systems inherently face include substantial intra-class variations and minimal inter-class differences [3]. FER is highly sensitive to intra-class variations, such as differences in age, gender, illumination, and facial pose [4]. Additionally, the small inter-class differences hinder the system's ability to accurately differentiate between distinct emotions. To address these issues, we propose applying the self-attention mechanism in our EmoFormer network, which establishes connections among various facial regions, such as the mouth, eyes, and nose. EmoFormer directs attention to the areas most involved in conveying emotions by detecting facial landmarks. This method identifies the most

This research was supported by the National Science Foundation (CAREER Award HCC-2053498).

Corresponding author: Ramana Vinjamuri.
F, Safavi, K. Patel, and R. Vinjamuri are with the Department of Electrical and Computer Science, University of Maryland, Baltimore County, MD 21250, USA (e-mail:rvinjam1@umbc.edu)

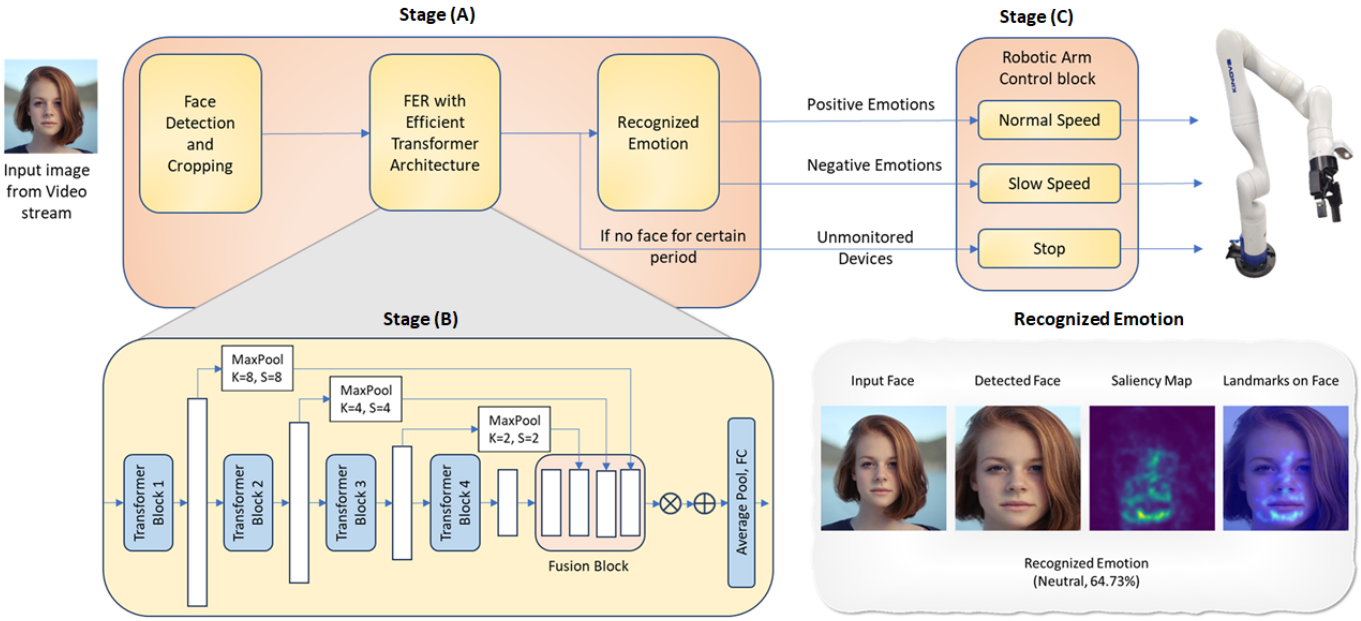

Fig. 2. This schematic illustrates the process flow of our HRI system, where user emotions directly influence the operational dynamics of a robotic arm. Key stages include: Stage (A) real-time facial recognition; Stage (B) emotion categorization via the advanced EmoFormer model; and Stage (C) responsive adjustment of the automated arm's movement speed based on detected emotional cues. Emotion recognition begins with Stage (A), which involves face detection and cropping. Following this, the EmoFormer segments the face and extracts facial landmarks in Stage (B), ultimately culminating in the application of emotion classification. The system changes behavior based on recognized emotions in Stage (C) emotional interaction. Additionally, the system halts the arm's function if no face is detected for a predefined duration in Stage (C).

significant regions for expression recognition. As observed in Figure 1, the blue dots represent facial landmarks detected by our new EmoFormer model. These landmarks are primarily located in facial areas crucial for emotion recognition, such as the mouth, nose, and eyes. Notably, they are not situated in less relevant areas for facial expressions, like the hair or jawline, focusing instead on the regions most indicative of emotions.

With the advent of deep learning models in recent years, Convolutional Neural Networks (CNNs) such as VGG [5] and ResNet [6] have served as backbones for FER systems. However, most of these architectures lack the integration of attention mechanisms, which are crucial for extracting more informative features from images in FER systems [3]. Although some models have incorporated pixel-level attention mechanisms, they often result in very large models, sacrificing efficiency for accuracy [7]–[9]. Additionally, these methods are not scalable, limiting their practical application, especially on mobile and computationally efficient embedded edge devices. To address these challenges, we have developed the EmoFormer series. This architecture is designed to adjust its size according to specific needs, balancing efficiency and accuracy while remaining adaptable for deployment in various scenarios.

With the popularity of Transformers [10] in natural language processing, there has been a recent surge in applying Transformers to computer vision tasks. Similar to transformer token embedding, facial images can be embedded into sequences of patches as visual words [11]. This approach allows for expression recognition from a global perspective. However, current FER methods based on Transformers are still inferior to those relying on CNN backbones in terms of classification accuracy, model size, and training cost [11]. While previous state-of-the-art methods primarily used attention and residual networks for FER, we explored the ability of the Mix Transformer for the first time in this domain. We leveraged the self-attention mechanism of the Mix Transformer, which can capture long-range relationships between patches. The main innovation, however, lies in the use of a fusion block in the EmoFormer architecture. The fusion block uses element-wise multiplication of facial landmarks, which are the outputs of the last layers, to emphasize the role of facial landmarks on the feature map. This fusion block assigns high importance to these facial landmarks in the transformed feature map produced by the Mix Transformer block. Additionally, the fusion block aggregates information from different layers, thereby integrating both local attention and global attention to render powerful representations.

The ensemble strategy for FER has been explored to enhance accuracy, with the network ensemble emerging as a notable approach. This technique entails ensembling individual networks by averaging their output predictions [12]. For instance, the Residual Masking Network [7], combined with seven other CNNs using a non-weighted sum average ensemble, achieved state-of-the-art accuracy for the FER2013 dataset [13]. Hence, we utilize the ensemble method in our research to boost accuracy and to achieve state-of-the-art (SOTA) results.

We conducted a case study by deploying our new model, EmoFormer, in the emotionally intelligent robotic arm. Our FER model captures the user's emotions and uses this information to control the speed of an automated arm's movement. As shown in Figure 2, our robotic arm incorporates affective

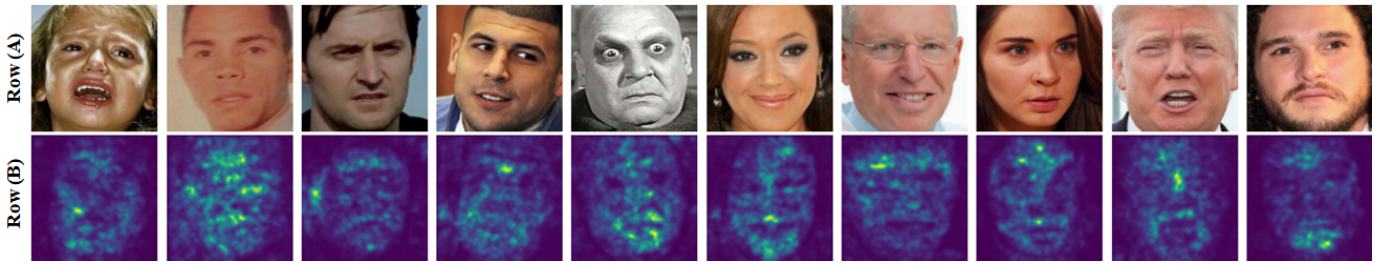

Fig. 3. Row (A) presents the original faces; Row (B) emphasizes the intricate pixel-level facial segmentation achieved by EmoFormer, utilizing the MiT Segmentation block on AffectNet dataset. It highlights the key facial areas identified by our EmoFormer model for emotion classification.

computing into our framework. The affective perception system in the robotic arm begins with face detection and cropping, followed by sending the cropped faces to EmoFormer, a transformer-based architecture, for emotion recognition. Emotions are classified as either positive or negative. If a positive emotion, such as a happy face, is detected, the robotic arm speeds up. Conversely, if a negative emotion, like a sad face, is perceived, the arm slows down. If the system doesn't detect a face or see any movement for a certain period, the robot stops operating. Additionally, our framework supports multimodal emotion recognition and can incorporate other modalities to control the robotic arm. Overall, the main contributions of our work can be outlined as follows:

1) We propose the robust and scalable EmoFormer series (B0 to B2), validated across both categorical and dimensional FER datasets. To the best of our knowledge, we are the first to apply Mix Transformer [14] for FER and incorporate a novel fusion block.
2) We extensively analyzed our architecture's accuracy and efficiency, including computational complexity, latency, and the number of parameters. EmoFormer-B0, our lightweight model, substantially outperforms heavier methods in efficiency while maintaining similar accuracy.
3) We validate EmoFormer series networks on DEAP [15], FER2013, and AffectNet, which show competitive performance with state-of-the-art benchmarks. Our EmoFormer ensemble with CNN methods achieved state-of-the-art results on FER2013 [13], AffectNet-7, and AffectNet-8 [16].
4) We conducted a case study using FER for the affective HRI system and validated emotionally intelligent human-robot interaction through a user experiment with six subjects.

The remainder of this article is organized as follows: Section II covers related work. Section III provides a detailed overview of our proposed methodology and delves into the underlying scientific principles of our approach. In Section IV, we describe the experimental setup, discuss the datasets used, and outline the accuracy and efficiency metrics implemented in our method. Section V discusses the results and offers further analysis. The article concludes in Section VI.

II. Related Work

A. Facial Expression Recognition

Facial expression recognition has been a classic problem in affective computing for decades. To tackle this problem, researchers have traditionally employed a variety of handcrafted features, including Local Binary Patterns (LBPs) [17] and Histogram of Oriented Gradients (HOGs) [18], among others [19], [20], to develop facial expression feature representations. In recent years, researchers have turned their attention to deep features for accurately extracting discriminative features [5], [6], [21]. Extracting features and achieving significant improvements in CNNs have gradually progressed over the last decade [22]. Researchers have investigated CNN models because these models capture not only low-level texture details but also high-level abstract representations from facial imagery. They discovered that incorporating attention mechanisms into deep CNNs can extract more informative features from images [3]. Consequently, many FER methods use a global self-attention mechanism with CNNs to identify facial expressions and highlight discriminative regions [23], [24]. The intuition behind using the attention mechanism stems from the fact that the essence of facial expressions lies in key regions of the face.

Facial expressions predominantly depend on certain facial regions, such as the eyes and mouth. On the other hand, regions like the hair and jawline play a minimal role in conveying emotions [7]. Certain architectures achieve pixel-level facial landmark detection by integrating segmentation modules within their structure. The Residual Masking Network [7], [8] boosts CNNs in facial expression detection by creating segmentation masks that highlight the most informative facial regions. These UNet-style masks are embedded within the layers of residual networks, enhancing network performance by increasing pixel-level attention to the facial landmarks. Therefore, using segmentation blocks to enhance pixel-level landmark masks for facial expressions can lead to increased model accuracy. After conducting extensive research on semantic segmentation [25]–[27], we integrated a 'Mix Transformer' block into our EmoFormer model. MiT blocks, derived from the SegFormer [14] architecture, serve as our segmentation module. As depicted in Figure 3, facial segmentation is clearly visible in the second row of the images. This segmentation block plays a crucial role in detecting pixel-level landmarks, which is essential for the classification processes in facial emotion recognition. Specifically, self-attention mechanism

within the MiT blocks enables the network to concentrate on the most pertinent facial regions, improving the accuracy of FER model.

Many studies show that feature fusion can strengthen the representational and generalization capabilities of entire networks, thereby boosting recognition performance. Therefore, various feature fusion methods have been proposed to further enhance expression performance [28]–[30]. The primary aim of fusion methods is to combine these feature maps through concatenation. They squeeze unnecessary information and produce correlated weight maps from both local and global perspectives [11], [31]. For instance, CBAM [32], enhances accuracy by integrating channel attention and spatial attention mechanisms, capturing richer information through a unified framework. As a result, we leverage a novel fusion approach in our EmoFormer to improve the network's accuracy.

There has been much less work investigating scalability and lightweight networks for FER. Most researches primarily applying established CNN architectures for use on mobile devices [33], [34]. Some researchers have explored a method that uses a lightweight local-feature extractor and a channel-spatial modulator with depthwise convolution to improve accuracy [35]. To achieve a simpler and more powerful architecture for FER tasks, POSTER++ uses a window-based cross-attention mechanism and prune the network by removing extra branches in the two-stream design [36]. However, these methods often involve trade-offs between accuracy and efficiency, making them less scalable across different scenarios. This is particularly challenging when considering the use of FER model for both high computational platforms and resource-limited systems

B. Transformers in Facial Expression Recognition

In recent years, the remarkable performance of transformers in natural language processing has catalyzed their application in other domains. Notably, the Vision Transformer (ViT) [37] pioneered the use of transformers in computer vision. The ViT framework exhibits excellent performance in classification tasks by segmenting each image into a sequence of patches, which are subsequently processed through multiple transformer layers for classification. With the rising popularity of ViT, numerous FER systems incorporating transformers have emerged. ViT-FER, introduced as the first ViT model for FER, focuses on learning local representations with relational awareness [38]. It utilizes Vision Transformers (ViT) to establish complex relationships between different local patches by incorporating both local and global scopes in representation learning. This approach enhances FER performance. Moreover, the multi-head self-attention mechanism in ViT allows simultaneous attention to features from various information subspaces at different positions, fostering relationships among diverse local patches.

The utilization of transformers for FER excels by incorporating a selective fusion block [11]. This transformer-based architecture effectively handles recognition tasks by capturing long dependencies between input sequences through the global self-attention mechanism. This global self-attention enables the model to overlook information-deficient regions and recognize expressions from a global perspective, even in cases of occlusions or varying poses. Transformers have also been used such that multiple non-overlapping attention regions extract data from different parts of faces [39]. The dual-direction attention models in transformers identify long-range dependencies, allowing for the capture of holistic and contextual facial information [3]. Consequently, in our work, the self-attention mechanism in MiT proved empirically effective in learning global information from facial images.

C. Affective Human-Robot Interaction

The primary objective of an emotionally intelligent robot is to enhance the quality of human-robot interaction (HRI) [40]. Numerous industrial informatics studies have incorporated emotional states for intelligent interaction-based industrial systems. For instance, a specific research study utilizes detected facial expressions as a direct feedback control mechanism in a learning control strategy for air conditioning to mitigate human sleepiness [41]. Additionally, Electroencephalogram (EEG)-based fatigue detection emphasizes the integration of emotions into robotic interfaces, offering a solution to enhance safety measures within the transportation industry [42]. Simultaneously, the evolution of social robotics has resulted in the development of robots adept at interpreting human emotions through social cues. For example, the social robot Ryan employs artificial emotional intelligence to aid older adults with depression and dementia [43]. By recognizing facial expressions and other cues, it discerns user emotions and responds with affective dialogues. Consequently, robots that can perceive and understand basic human emotions enhance interactions within human environments, boosting operational effectiveness and productivity in industry. Facial expressions are a primary medium for conveying emotions during interactions, reflecting users' attitudes and responses in HRI. For seamless and prompt communication, a reliable and precise FER technique is essential.

III. Methodology

In the following sections, we describe the various components of our affective HRI system, illustrated in Figure 2. This system enables real-time adjustment of movements based on detected user facial expressions. We examine the following subsections, which explain the architecture and all methods used in this experiment: EmoFormer Architecture, EmoFormer for Dimensional Classification, Ensemble Method, and Affective Robotic Arm.

A. EmoFormer Architecture

As illustrated in Figure 4, we outline the structure of our novel EmoFormer model, designed for Facial Expression Recognition (FER) predictions. We have developed a series of EmoFormer models — EmoFormer-B0, EmoFormer-B1, and EmoFormer-B2 — all of which share the same basic architecture. These models differ in terms of their underlying MiT variants [14], resulting in variations in depth and computational complexity.

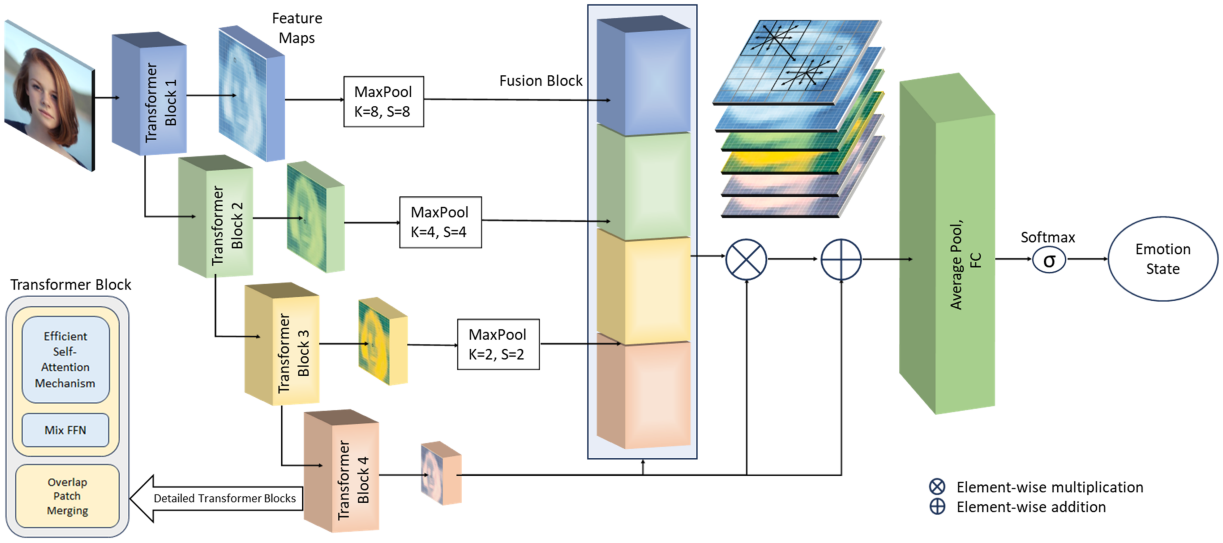

Fig. 4. Diagram of the EmoFormer model series, including EmoFormer B0, B1, and B2, each employing the same architecture featuring hierarchical MiT blocks and an advanced fusion mechanism. This design, incorporating Transformer blocks 1 through 4 and a fusion block, merges multiscale features, aiming to enhance accuracy in FER predictions.

Additionally, we introduce an enhanced fusion mechanism, as illustrated in the fusion block in Figure 4, that concatenates multiscale feature maps along the channel dimension. We then utilize the output of this fusion block to element-wise score the importance of the final output from the Transformer block 4. This augmented fusion mechanism combines high-resolution coarse features with low-resolution fine features, improving accuracy for classification tasks. In the following subsections, we describe the architecture of Mix Transformer and the Fusion Blocks.

1) Mix Transformer Blocks: Integrating the lightweight MiT, which functions as the encoder block within the SegFormer [14] semantic segmentation framework, is central to our design. Our MiT variants range from the lightweight B0 version to the more extended versions B1 and B2. These networks produce four hierarchical feature maps and are denoted as F_i. To explain the network flow, an input image with dimensions $H \times W \times 3$ is first divided into n patches of size 4×4 pixels ($Patch_{4\times4}$). These patches are then processed through Mix Transformer Blocks, which generate multi-level feature maps as outlined in Equation 1.

$$F_i = \text{MiT}(n \times Patch_{4\times4}) \quad \text{for } i = 1, 2, 3, 4 \qquad (1)$$

These feature maps capture a spectrum of features, from coarse to fine-grained, as shown in Figure 4. As illustrated in Table II, the transformer-based architecture generates four feature maps, and as a result, i spans the values in the set $\{1, 2, 3, 4\}$. The resolutions of the resulting feature maps are downscaled to $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{16}$, and $\frac{1}{32}$ of the original image size. For an input image with dimensions $224 \times 224 \times 3$, the MiT model produces hierarchical feature maps F_i with resolutions represented as $H_i \times W_i \times C_{i+1}$. As seen in the output size B1, B2 column in Table II, the output dimensions H_i and W_i take values from $\{56, 28, 14, 7\}$, inspired by the design principles of the ResNet architecture [6]. The number of output channels C_{i+1} is $\{64, 128, 256, 512\}$ for both B1 and B2. The primary reason for choosing these specific numbers is to follow the structure of ResNet, which offers advantages such as efficient feature extraction at different stages and a proven balance of performance and efficiency. As seen in the output size B0 column in Table II, the values $\{32, 64, 160, 256\}$ are chosen for B0 to cater to real-time applications due to its lower computational cost. The number of output channels for the subsequent Transformer Blocks, C_{i+1}, are always greater than the corresponding previous C_i values. This reflects the design principle of deep networks where increased depth is used to capture more complex features. As shown in the Detail column of Table II, we adhere to the MiT block [14] principles by defining Ker, S, and P, where Ker denotes kernal size, S is the stride between adjacent patches, and P indicates the padding size.

As depicted in Figure 4, each Transformer block comprises overlapping patch embedding and a Transformer Encoder, which contains the attention mechanism, attention heads, and the Multi-Layer Perceptron (MLP). As shown in Table I, the basic structure of the spatial reduction ratio in the attention mechanism (R), the number of attention heads (N), and the MLP expansion ratio (E) in the Transformer block are the same in the series of EmoFormer-B0, B1, and B2. However, the main difference lies in the number of layers and output channels within the Transformer blocks among these three models. B2 is deeper than the B0 and B1 models, as illustrated in Table I, and the difference between B0 and B1 is based on the number of output channels, shown in TABLE II.

The MiT blocks spatial dimensions are reduced for computational efficiency and to build a hierarchical representation of the input. Moreover, Overlapped Patch Merging, Efficient Self-Attention, and Mix-FFN [14], are essential components for enhancing the efficiency of the vision transformer and increasing the effective receptive fields. The Overlapped Patch Merging operation reduces the size of feature maps while pre-

TABLE I
COMPARISON OF THE NUMBER OF LAYERS IN EMOFORMER MODELS B0, B1, AND B2 ACROSS DIFFERENT STAGES.

Block	(R, N, E)	B0, B1 Layers	B2 Layers
1	(8, 1, 8)	2	3
2	(4, 2, 8)	2	3
3	(2, 5, 4)	2	6
4	(1, 8, 4)	2	3

serving their local continuity. Furthermore, it enables patches to merge into a compact vector. This is achieved through convolution to extract patches, flattening and transposing to organize patches into sequences, and normalization to stabilize and enhance the representation. This transformation allows for a more efficient and compact representation of the image data. The Self-Attention block reshapes and transforms a sequence of K elements into K', reorganizing and compressing the computational complexity associated with the self-attention mechanism, as outlined in Equation 2 [14].

$$\text{Attention}(Q, K', V) = \text{Softmax}\left(\frac{QK'^T}{\sqrt{d_{\text{head}}}}\right)V \tag{2}$$

Ultimately, Mix-FFN employs a combination of MLP (Multi-Layer Perceptron) and a 3×3 Convolutional layer in its feed-forward network rather than relying on positional embeddings [14].

The MiT [14] present a more compact alternative to conventional vision transformers (ViT) [44]. The efficient self-attention mechanism enables our proposed method to be especially well-suited for extracting facial representations, particularly for segmenting facial areas that determine emotions. Unlike the ViT, the proposed vision transformer omits positional embeddings, enhancing the efficiency of FER classification without sacrificing accuracy. Meanwhile, MiT is capable of capturing relationships between sequences of patches as visual words, providing global context information with attention to facial landmarks. Subsequently, we employ our novel fusion method to combine feature maps of different spatial sizes.

2) Fusion Block: This block combines the feature maps F_i to encompass fine features at low resolution and coarse features at high resolution. As depicted in Figure 4, we initially employ max pooling operations with varying kernel and stride sizes of 8, 4, and 2 to downsample the feature maps F_1, F_2, and F_3. Subsequently, these downsampled feature maps, along with F_4, are concatenated, and the channel count is reduced using a 1×1 convolutional layer, resulting in the creation of F_R. We anticipate that F_R will contribute to enhancing the accuracy of the feature maps. Consequently, we utilize the output of the fusion block F_R to element-wise score the importance of the final output of the Transformer blocks F_4 using the following operation, as outlined in Equation 3:

$$F_N = F_4 + F_4 \otimes F_R \tag{3}$$

This augmented fusion mechanism gives more weight to the features extracted by F4, which are facial landmarks. In this novel mechanism, we emphasize regions that have a greater impact on facial expressions. Additionally, FR contains both fine-grained and coarse features. These features are derived from four feature maps, which originate from transformer blocks. In the concluding step, F_N undergoes an average pooling operation followed by a fully connected layer. This layer includes a dropout component with a dropout rate of 0.4, as well as a linear transformation specifically designed to map the input to distinct output classes. For FER2013 and AffectNet-7, it translates to 7 unique classes, whereas AffectNet-8 consists of 8. The network employs an average pooling layer, followed by a fully connected layer with SoftMax activation. This final layer yields outputs corresponding to distinct facial expression states.

B. EmoFormer for Dimensional Classification

The EmoFormer method is used for the dimensional classification of arousal and valence per subject in the videos in the DEAP dataset. We used our EmoFormer as a facial feature extractor to process a sequence of facial images and extract meaningful features. The process begins with input images, which are preprocessed and individually passed through EmoFormer. EmoFormer processes each image to produce feature maps with reduced spatial dimensions.

Since we have a video per subject, we use a unique technique inspired by [45] to reduce the spatial dimensions and then handle the temporal sequence. The model processes the sequence of feature maps from EmoFormer to capture temporal dependencies across the sequence of images for higher accuracy [45]. The final output is a feature vector capturing the essential facial features from the input image sequence. This feature vector, represented as a $1 \times \mu_2$ vector, is denoted as $f_{\text{face}} = (E_1, E_2, \ldots, E_{\mu_2})$. Following the DeepVaNet architecture [15], the size of the feature vector is 16 in our experiment.

Finally, this feature vector is passed through a multi-layer classifier that sequentially reduces the dimensionality and applies activation functions, culminating in a sigmoid function to produce the final output. In this approach, we use EmoFormer as the facial feature extractor for the dimensional classification of emotions.

C. Ensemble Method

Ensemble methods significantly enhance prediction accuracy by integrating multiple models. In our study, we illustrate this by integrating EmoFormer with various CNNs and transformer-based networks through a straightforward non-weighted average ensemble. This method amalgamates the predictive outcomes of different CNNs and deep transformer networks, relying on an equal contribution from each model. Generating ensemble results involves deploying multiple models and merging their predictions using a weighted averaging approach [7]. Let $M = \{m_1, m_2, \ldots, m_n\}$ represent the set of models, each producing a result vector r_i. The aggregated results vector R' is computed as:

$$R' = \sum_{i=1}^{n} p_i \cdot r_i \tag{4}$$

TABLE II
CONFIGURATION SETTINGS OF EMOFORMER SERIES

Layer name	Output Size	Output Size B0		Output Size B1, B2		Detail
Transformer Block 1	$F_1 = C_1 \times \frac{H}{4} \times \frac{W}{4}$	C1 = 32	32 × 56 × 56	C1 = 64	64 × 56 × 56	Ker = 7, S = 4, P = 3
Transformer Block 2	$F_2 = C_2 \times \frac{H}{8} \times \frac{W}{8}$	C2 = 64	64 × 28 × 28	C2 = 128	128 × 28 × 28	Ker = 3, S = 2, P = 1
Transformer Block 3	$F_3 = C_3 \times \frac{H}{16} \times \frac{W}{16}$	C3 = 160	160 × 14 × 14	C3 = 256	256 × 14 × 14	Ker = 3, S = 2, P = 1
Transformer Block 4	$F_4 = C_4 \times \frac{H}{32} \times \frac{W}{32}$	C4 = 256	256 × 7 × 7	C4 = 512	512 × 7 × 7	Ker = 3, S = 2, P = 1
Fusion Block	$F_R = C_4 \times \frac{H}{32} \times \frac{W}{32}$	C4 = 256	256 × 7 × 7	C4 = 512	512 × 7 × 7	MaxPool2, Concat
Average pooling	$C_4 \times 1 \times 1$	C4 = 256	256 × 1 × 1	C4 = 512	512 × 1 × 1	Adaptive Average Pooling
FC, Softmax	number of classes = 7 for FER2013 and AffectNet-7, 8 for AffectNet-8					Dropout (p=0.4)

where $p_i = 1$ for all i in the given script. Given the true targets $T = \{t_1, t_2, \ldots, t_k\}$, the predicted class for each test instance is:

$$C_j = \arg\max(R'_j), \quad \text{for } j = 1, 2, \ldots, k \tag{5}$$

Finally, accuracy A is calculated as:

$$A = \frac{1}{k} \sum_{j=1}^{k} \mathbb{I}(C_j = T_j) \times 100 \tag{6}$$

Subsequently, we assess the ensemble's accuracy by comparing its predictions to the ground truth labels. The goal is to evaluate the performance of various models within an ensemble configuration and investigate the ensemble model's impact on accuracy, aiming to achieve state-of-the-art results. where $\mathbb{I}$ is the indicator function, and k is the total number of test instances.

D. Affective Robotic Arm

In the case study of EmoFormer, we developed an affective robotic arm that responds and adapts its actions based on the user's facial expressions. Similar to many other robots, our robotic arm is equipped with a camera that captures images. This integrated camera can capture images, which can then be passed as a sequence into EmoFormer for facial expression recognition. By integrating FER with robotic control, we created an emotionally responsive human-robot interaction system. This setup is used to evaluate the user experience by allowing human facial expressions and emotions to directly influence the robot's behavior. We employed the Kinova Gen3 Lite robotic arm as the primary platform in our research study, seamlessly integrating the emotion recognition system into its control framework.

We implemented a real-time speed adjustment mechanism that dynamically modulates the arm's movements in response to the recognized emotions in the user. This mechanism aims to integrate affective computing technologies to enhance human-robot interactions and evaluate user comfort and overall satisfaction during collaborative interactions. We use EmoFormer for facial expression recognition and inferring emotional modes. As provided in Algorithm 1, the most frequently observed expression within a rolling time window determines the arm's movement speed. Before processing emotions, the system first checks for the presence of a face. If no face is detected, it ceases operation. Assuming a face is detected, we assigned distinct speed settings for positive and negative emotions to investigate their impact on the user's perception and experience. For positive emotions, including happiness and neutrality, the robotic arm operated at a speed of 15 revolutions per minute (rpm). We hypothesized that this faster arm movement would align with the user's positive emotional state. Fast movements heightened sense of responsiveness and engagement [46]. In contrast, for negative emotions including sadness, fear, anger, and disgust, the robotic arm operated at a reduced speed of 5 rpm. Slower movements, potentially alleviating any discomfort or unease associated with negative emotional states and promoting a calming and reassuring interaction [47]. This approach represents a novel and innovative exploration of emotion-aware robotic manipulation. It demonstrating its potential to enhance human-robot collaboration across various emotional contexts.

Algorithm 1 Affective Robotic Arm

Require: User's Facial Expression
Result: Kinova Gen 3 robotic arm movement speed

```
Determine the user's expression by EmoFormer
Initialize rolling time window for emotion observation
Determine the most frequently observed emotion in the window
if no face detected then
    Cease the operation
end if
if emotion ∈ {happiness, neutrality} then
    Set robotic arm speed to 15 rpm
    Expect heightened responsiveness and engagement
else if emotion ∈ {sadness, fear, anger, disgust} then
    Set robotic arm speed to 5 rpm
    Expect to provide calming and reassuring interaction
end if
```

E. Experiment Setup

IV. EXPERIMENT

We conducted extensive experiments on four frequently used datasets: FER2013 [13], AffectNet-7, AffectNet-8 [16], and DEAP [15]. These datasets provide a variety of scenarios to verify the generalizability and robustness of our method. For instance, the AffectNet dataset is collected in the wild, which may suffer from different illuminations and occlusions. We compared the new EmoFormer model in terms of accuracy and efficiency and performed a thorough analysis. In this section,

we describe the details of the datasets used in the experiments, the efficiency and accuracy metrics, and the experimental setup for our work.

A. FER2013

In this study, we employed the FER2013 dataset [13] as a reference dataset to assess the performance of various FER models in terms of accuracy. This dataset consists of 33,572 grayscale facial images with dimensions of 48x48 pixels. The dataset is divided into seven standard categories: Angry (4,953 images), Disgust (547 images), Fear (5,121 images), Happy (8,989 images), Sad (6,077 images), Surprise (4,002 images), and Neutral (6,198 images). Notably, FER2013 achieves a level of accuracy comparable to that of humans, approximately 65±5%, and the most effective algorithm achieves an accuracy rate of 76.82% in correctly identifying facial expressions [7].

B. AffectNet

To assess the efficacy of our proposed model across various datasets and its adaptability to real-world scenarios for facial emotion recognition, we utilized the AffectNet dataset [16] for training. The comprehensive nature of the AffectNet dataset, which encompasses diverse facial expressions in authentic environments, is noteworthy. The dataset comprises images manually labeled into eight distinct emotional states, known as AffectNet-8, which include neutral, happy, angry, sad, fear, surprise, disgust, and contempt. The seven expression categories, referred to as AffectNet-7, include the same expressions except for contempt. Our training process utilized 287,657 images sourced from the AffectNet dataset. For evaluation, we employed the official test set containing 4,000 images, evenly distributed with 500 images per emotional category.

C. DEAP

Database for emotion Analysis using physiological signals dataset is a widely used multimodal dataset designed to analyze human emotional states. This dataset comprises frontal face videos of 22 subjects while they watched 40 music videos, each selected for its potential to elicit a range of emotional responses. For emotion classification, the DEAP dataset provides labels for valence and arousal, with participants rating their emotions on a discrete 9-point scale for valence and arousal. It enables the categorization of emotional states into distinct classes. In addition to video recordings of face, physiological data was collected from 32 participants (16 males and 16 females), providing a comprehensive dataset for multimodal emotion analysis. The DEAP dataset offers a detailed analysis based on facial expressions captured in real-time and enables comparisons of FER performance with other modalities such as electroencephalography (EEG) signals.

D. Implementation Details

For FER2013, the training images were resized to 224×224 pixels and data augmentation methods included flipping and rotation. The training was conducted for 50 epochs with Cross-Entropy Loss as the loss function. The learning rate was set to 10^{-4}.

To assess our model's generalizability, we also tested our models on the AffectNet dataset. In the AffectNet experiments, images were preprocessed and resized to 224×224 pixels. We utilized the Adam optimizer, adjusting training parameters over eight epochs. Initially, for the first three epochs, the learning rate was set to 10^{-3}, focusing exclusively on the weights of the last layer of the EmoFormer. In the subsequent five epochs, the entire network was trained with a learning rate of 10^{-4}.

For DEAP dataset, images were preprocessed and resized to 224×224 pixels. We use binary cross-entropy as the loss function. During inference, we pass the test images through our proposed network to obtain a FER score. Based on this score, the final binary prediction is made for arousal and valance: if the score is greater than 0.5, the prediction is "High"; otherwise, it is "Low". We train and test our model on each individual subject for 22 subjects, a process referred to as a per-subject experiment. Our model undergoes 10-fold cross-validation, and the average testing accuracy is used to measure performance. For validation, we utilize the mean recognition accuracy of both valence and arousal. All EmoFormer models are trained on a single system equipped with an NVIDIA GeForce RTX 3090 GPU and an Intel Core i9 processor.

E. Accuracy Evaluation

To gauge the accuracy of our classification models, we employ the following formula [7]:

$$\text{Accuracy (Acc)} = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

Here, the correctly predicted pixels are denoted as true positives (TP), while those correctly identified as not belonging to a specific class are referred to as true negatives (TN). Pixels that belong to the category but are incorrectly predicted as a different type are categorized as false negatives (FN), and finally, the pixels mistakenly indicated as belonging to the class are termed false positives (FP).

F. Efficiency Metrics

To evaluate the efficiency of our models, we consider three key metrics: the number of trainable parameters, computational complexity (Flops), and inference time (as outlined in TABLE I).

1) Learnable Parameters (Parm): This metric quantifies the complexity of a model by counting the total number of learnable parameters within a feed-forward neural network.

2) FLOPs (Floating-Point Operations): FLOPs measure the total number of calculations required to complete a single forward pass.

3) Inference Time (Time): Inference time is calculated on a single RTX 3090 GPU using CUDA 11.7 and PyTorch 1.13.0. After initializing the GPU with dummy examples, the network is executed 300 times with an input resolution of 224 × 224 and a batch size of 48. The resulting average time is then

TABLE III
PERFORMANCE EVALUATION OF NETWORKS ON FER2013

	Models	FLOPs	Params	Time	Acc(%)
Efficient	MobileNetV3	0.23B	5.48M	7.48ns	64.78
	MobiExpressNet	1.08M	14.4M	-	67.96
	Imp-MobileNetV3	0.19B	1.29M	11.68ns	68.14
	RASN	1.82B	75.08K	-	71.44
	EmoFormer-B0	**0.425B**	**3.45M**	**7.74ns**	**73.47**
Non Efficient	Ad-Corre	4.55B	21.38M	-	72.63
	ResAttNet56	6.33B	29.77M	17.69ns	72.63
	Densenet121	2.89B	6.96M	19.57ns	73.16
	Resnet152	11.60B	58.16M	23.66ns	73.22
	Cbam_resnet50	4.14B	26.05M	16.94ns	73.39
	ResMaskingNet	26.76B	142.9M	17.63ns	74.14
	LHC-Net	-	32.4M	-	74.42
	EmoFormer-B1	**1.63B**	**13.67M**	**7.43ns**	**74.14**
	EmoFormer-B2	**3.17B**	**24.72M**	**13.94ns**	**74.48**
	Ensemble Model	-			**77.35**

reported. For real-time model consideration, the standard video streaming rate is set at 24 frames per second (fps), meaning that if a model processes an image in less than 41ms, it qualifies as a real-time model.

V. Results and Discussion

A. Accuracy and Efficiency of EmoFormer

To evaluate our method's performance, we initially conducted a two-fold analysis of the FER2013 public dataset to ascertain the method's accuracy and efficiency. We then extended the network classification to include three datasets: FER2013, AffectNet, and DEAP. The results for all these experiments are provided in Tables III, IV, and V. Firstly, we concentrated on networks known for their high accuracy on FER2013, which included Ad-corre [48], ResmaskingNet [7], Resnet151 [49], Densenet121 [50], RessAttNet56 [51], Cbam_resnet50 [52] and LHC-Net [53]. Secondly, we examined recent efficient networks, specifically MobileNetV3 [54], MobiExpressNet [55], Improved MobileNetV3 (imp-MobileNetV3) [56], and RASN [57]. As illustrated in Table III, within the non-efficient group of methods evaluated on the FER2013 dataset, our approach EmoFormer-B0 consistently demonstrated the highest efficiency across all efficiency metrics, including FLOPs, the number of learnable parameters, and inference time. Moreover, among the lightweight models listed in the upper section of Table III, EmoFormer-B0 achieved the highest accuracy while maintaining comparable efficiency. EmoFormer inference times are significantly faster than the requirements for video streaming, making them suitable for FER in video streaming.

By effectively utilizing transformer blocks in a compact hierarchical design and excluding positional embeddings from the original vision transformer structure, following the SegFormer architecture [14], our EmoFormer-B0 model was able to achieve a low FLOPs value of 425 million as provided in Table III. This resulted in high accuracy with minimal computational overhead, putting it on par with lightweight methods like MobileNetV3 [54]. Similarly, our EmoFormer-B0 deep network excelled in terms of efficiency, with an efficiency score of only 3.45 learnable parameters, which was

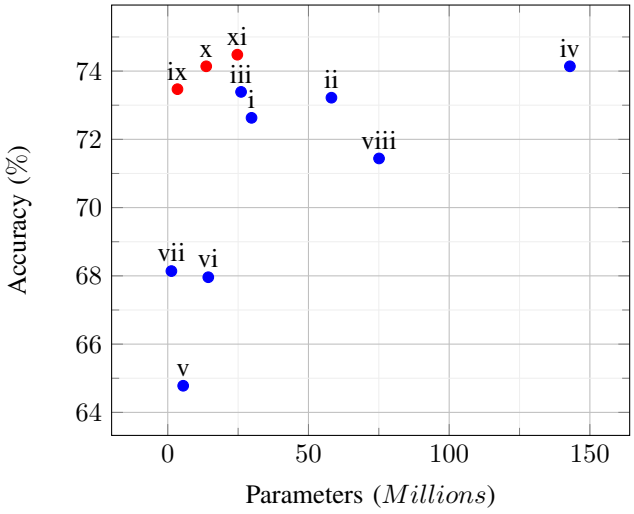

Fig. 5. Plot of accuracy vs number of learnable parameters. The different networks are represented as follows: i) ResAttNet56, ii) Resnet152, iii) Cbam-resnet50, iv) ResMaskingNet, v) MobileNetV3, vi) MobiExpressNet, vii) Imp-MobileNetV3, viii) RASN, ix) EmoFomer-B0, x) EmoFormer-B1, xi) EmoFormer-B2.

TABLE IV
PERFORMANCE EVALUATION OF NETWORKS ON AffectNet 8-CLASS AND AffectNet 7-CLASS

Model	FLOPs	Params	AffectNet-8 (%)	AffectNet-7 (%)
Poster++	8.4B	43.7M	63.77	67.49
DDAM	0.5B	4.0M	64.25	67.03
DAN	2.2B	19.0M	62.09	65.69
MA-Net	3.65B	50.54M	60.29	64.53
EfficientFace	0.154B	1.28M	59.89	63.70
EmoFormer-B0	0.425B	3.45M	60.91	64.51
EmoFormer-B1	1.63B	13.67M	61.51	64.62
EmoFormer-B2	3.17B	24.72M	62.01	65.48
Ensemble Model	-		**65.14**	**67.71**

41 times lower than that of ResMaskingNet and comparable to contemporary lightweight models like RASN [57]. By adopting a compact model with fewer parameters, our proposed method reduced memory requirements and computational burden. Furthermore, our model exhibited impressive inference speed, with an average inference time of 7.74 nanoseconds (ns), making it well-suited for real-time applications, akin to lightweight approaches such as MobileNetV3 and Imp-MobileNetV3 [56].

In addition to its efficiency, our proposed networks achieved high accuracy rates: EmoFormer-B0 at 73.47%, EmoFormer-B1 at 74.14%, and EmoFormer-B2 at 74.48%. Furthermore, an ensemble model, comprising the MiT-based models and supplemented with ResMasking, BAM_ResNet50, and ResNet152, achieved state-of-the-art performance on the FER2013 dataset with an accuracy of 77.35%. Figure 5 presents a plot of accuracy versus the number of learnable parameters, serving as one of the metrics of efficiency. In the plot, our models are marked in red, while all other models are marked in blue. It is evident from Figure 5 that our models demonstrate efficiency by maintaining high accuracy with fewer learnable parameters in comparison to the other models.

Another notable dataset used to evaluate our model's performance is the AffectNet dataset. The results are shown in

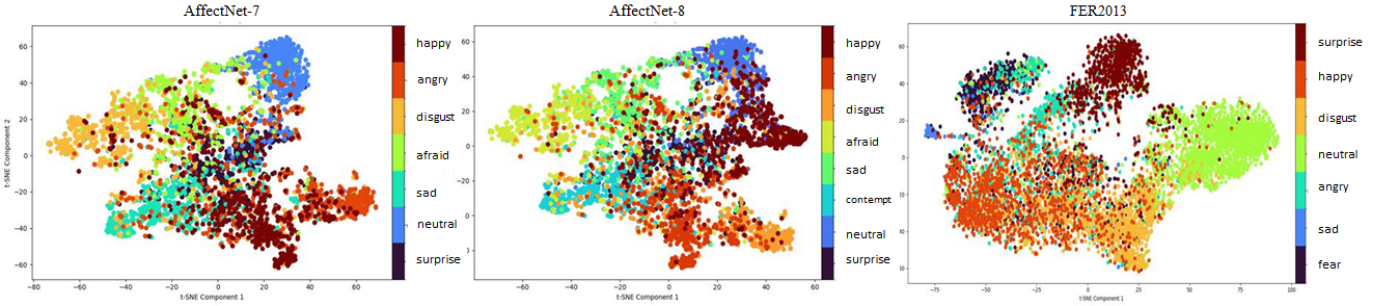

Fig. 6. The graphs represent 2D t-SNE visualizations of the extracted high-dimensional features of EmoFormer-B2 on the AffectNet-7, AffectNet-8, and FER2013 datasets.

TABLE V
ACCURACY OF EEG AND FUSION(EEG + FACE) BASED MODELS FOR VALENCE AND AROUSAL DETECTION ON DEAP DATASET

Models	Valence	Arousal
EEG		
DCNN + ConvLSTM	87.84	87.69
DCNN	90.62	86.13
Deep Residual Network	97.75	99.03
EEG + Face		
3D-CNN	96.13	96.79
CNN	96.69	97.15
CNN + Bi-LSTM	95.03	94.94
(Our model) Face		
EmoFormer-B0	97.86	98.07

TABLE VI
ACCURACY OF EMOTIONS USING EMOFORMER TRAINED ON FER2013

Models	Ang	Dis	Fear	Hap	Sad	Sur	Neu
EmoFormer-B0	63.1	76.4	58.5	90.2	61.8	84.4	74.3
EmoFormer-B1	64.8	72.7	59.7	90.6	63.6	83.4	74.6
EmoFormer-B2	67.2	80.0	57.6	91.5	61.4	85.8	74.9

Table IV, where we compare our model with some of the most recent models, namely Poster++ [36], DDAM [58], DAN [59], MA-Net [60], and EfficientFace [61]. EmoFormer-B0 is comparable to EfficientFace and has the second lowest FLOPs and parameters, with 0.425B and 3.45M, respectively.

The accuracies of EmoFormer-B0, EmoFormer-B1, and EmoFormer-B2 consistently increase as the number of parameters increases. Despite having almost 14 times fewer parameters, EmoFormer-B0 comfortably outperforms MA-Net [60], highlighting the efficiency of our model. The accuracies of EmoFormer-B0, EmoFormer-B1, and EmoFormer-B2 are 60.91%, 61.51%, and 62.01%, respectively, on AffectNet-8. Similarly, EmoFormer-B0, EmoFormer-B1, and EmoFormer-B2 achieved accuracies of 64.51%, 64.62%, and 65.48% on the AffectNet-7 task, respectively. We achieved new state-of-the-art accuracy of 65.14% on the AffectNet-8 through an ensemble model comprising of EmoFormer-B0, EmoFormer-B1, EmoFormer-B2, complimented with DDAM, Poster++, EfficientNet-B0 and EfficientNet-B2 [62]. Furthermore, we also achieved a cutting edge accuracy of 67.71% on the AffectNet-7, the ensemble models including EmoFormer-B0, EmoFormer-B1, EmoFormer-B2, and Poster++. These results underscore the importance of integrating Transformers for higher accuracy in FER.

TABLE VII
ACCURACY OF EMOTIONS USING EMOFORMER ON AFFECTNET.

Models	Neu	Hap	Sad	Sur	Fear	Dis	Ang	Cont
EmoFormer-B0	53.6	69.6	59.2	57.6	68.0	61.2	54.0	64.13
EmoFormer-B1	53.4	69.8	64.2	67.2	61.4	58.0	58.8	59.32
EmoFormer-B2	57.8	73.8	63.6	63.6	57.8	60.6	58.0	63.32

Moreover, our models exhibited strong performance on the DEAP dataset, a widely used dataset for the detection of human emotion states, which contains video facial recordings and neurophysiological signals. We applied the EmoFormer-B0 model to the facial features to predict valence and arousal. Valence represents the rating of how good or bad an emotion is, while arousal indicates the intensity of the emotion. Our most efficient model, EmoFormer-B0, achieved accuracies of 97.86% and 98.07% on valence and arousal affective states, respectively. This significant achievement is highlighted in Table V, which showcases various models that consider only neurophysiological signals like DCNN + ConvLSTM [63], DCNN [64] and Deep Residual Network [65] and the fusion models combining Neurophysiological and Facial features like 3D-CNN [66], CNN [67], CNN + Bi-LSTM [68]. Transfer learning was employed from the FER2013 EmoFormer-B0 while training on the DEAP dataset. It is noteworthy that EmoFormer-B0 outperforms the state-of-the-art fusion and EEG-based models, highlighting the model's versatility and its strong capability to be adapted for emotional analysis on diverse image datasets.

This superior accuracy can be attributed to our hierarchical Transformer Blocks with a larger effective receptive field and our novel fusion block, striking a balance between efficiency and accuracy. Our model distinguished itself for its efficiency in terms of FLOPs, the number of parameters, and inference time, offering low complexity, a compact footprint, and swift inference while maintaining competitive accuracy on all three datasets—FER2013, AffectNet, and DEAP.

B. Granular Emotion Performance Analysis

Table VI demonstrates the performance of three different models (EmoFormer-B0, B1, and B2) on the FER dataset across seven distinct facial expressions: Anger (Ang), Disgust (Dis), Fear, Happiness (Hap), Sadness (Sad), Surprise (Sur), and Neutral (Neu). Table VII extends the evaluation to the AffectNet dataset, covering eight facial expressions, including

Contempt (Cont). Similar to the FER dataset, the accuracy varies across expressions and models.

As observed in Figure 6, Table VI, and Table VII, there is a direct correlation between the 2D t-SNE visualization and the accuracies associated with recognizing facial expressions. Categories with well-defined clusters tend to have higher accuracies. For instance, the "Happiness" cluster in the 2D t-SNE visualization is well-defined, resulting in the highest accuracy of 91.5% on the FER2013 dataset and 73.8% on the AffectNet-8 dataset. In contrast, the "Fear" cluster is not clearly defined, with data points scattered throughout the plot, leading to the lowest accuracies of 57.6% on FER2013 and 57.8% on the AffectNet-8 dataset.

The models tend to have higher accuracy in recognizing certain expressions. The visibility and distinctiveness of facial features contribute to the varying accuracy in emotion recognition across different emotions. Happiness is the emotion that yields the best results across multiple datasets. The superiority of EmoFormer for happiness has shown 91.5% accuracy on FER2013 and 73.8% on AffectNet. Many different studies confirm that even when using various FER datasets and models, happiness consistently produces the best results [3], [7], [11], [36] . This is because happiness is associated with clear and distinctive facial movements, such as wide smiles, visibility of teeth, and prominent mouth and cheek expressions, making it easier for models to recognize.

Similarly, the superior accuracy of surprise in our FER2013 dataset, with the second highest accuracy of 85.8%, is associated with features such as an open mouth and raised eyebrows, which clearly indicate surprise and are confirmed in other studies [3], [7], [11]. On the other hand, certain emotions, such as fear, have comparatively lower accuracy rates. Fear involves more subtle and less uniform facial expressions, which vary more between individuals and are more challenging for recognition systems to accurately capture and classify. Consequently, improving the recognition of these emotions could significantly enhance the overall model performance.

The EmoFormer-B2 model often outperforms the EmoFormer-B0 and EmoFormer-B1 models in most categories. When comparing the performance, the EmoFormer-B2 model, which has a deeper architecture than B0 and B1, consistently shows better performance across datasets. This indicates that sacrificing efficiency for better accuracy is a major issue in deep networks. Deeper networks provide better results on large datasets and are more generalizable. Therefore, developing scalable methods such as EmoFormer is crucial for adaptability, especially for resource-limited platforms.

C. Mix Transformer vs. Vision Transformer: Landmark Detection Analysis

In this section, we evaluate the performance of the MiT model over the ViT model in facial expression recognition, leveraging saliency maps for deeper insight. After conducting a comprehensive analysis of various real-time semantic segmentation models [25], [26], we identified MiT [14] blocks as an ideal choice for extracting facial expression features, functioning similarly to segmentation masks in FER systems [7]. MiT blocks contain efficient self-attention blocks, a characteristic that enables the network to capture global facial semantic relational representations. Moreover, these efficient self-attention blocks effectively address computational complexity issues. Additionally, the hierarchical structure of MiT blocks adeptly extracts fine-grained and coarse features to capture both local and global attributes of the face. As demonstrated in Figure 7, our analysis reveals the superior effectiveness of MIT's architecture compared to ViT in identifying facial features. This underlines the importance of employing the MiT for enhanced facial expression recognition.

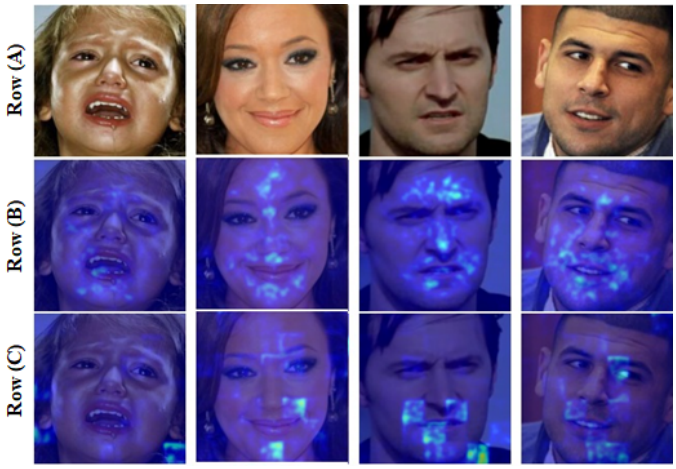

Fig. 7. Row (A) displays the original images from the AffectNet dataset, followed by Row (B) showing the saliency map from the MiT-based model, where EmoFormer identifies light blue dots around critical facial regions such as the eyes, nose, eyebrows, and mouth. Row (C) shows the saliency map from the ViT model, which lacks the distinctive light blue dots seen in Row (B).

The relationship between facial features and the expression of emotions is well-documented in scientific literature [69]. The Facial Action Coding System (FACS) describes all visually distinguishable facial activities based on 44 unique action units (AUs), each associated with particular emotions. For example, the Inner Brow Raiser (Frontalis, Pars Medialis) is associated with expressions of surprise and concern, while the Lip Corner Puller (Zygomatic Major) is typically seen in expressions of happiness or amusement [69]. These specific action units highlight the importance of critical facial regions such as the eyes, eyebrows, and mouth in the anatomy and physiology of the face in expressing emotions. As depicted in Figure 6, in the MiT-based model, light blue dots are distinctly visible around the critical facial regions in the middle row of the image, while these action units are much less identified in the ViT model. These visual observations are consistent with the actual accuracy metrics, as the MiT-based model consistently outperforms the ViT-based model. This is further underscored in Table II, which compares the MiT to the ViT-BASE model. This demonstrates that focusing on critical facial regions, identified by specific action units, enhances the accuracy of emotion recognition models. It is worth noting that in MiT-based models, the input image is segmented into smaller patches, each measuring 4x4 pixels, compared to the larger patch sizes of 16 or 32 pixels used in ViT models. This smaller patch size might be a contributing factor to the

TABLE VIII
QUESTIONS FOR USER RESPONSES ON ROBOTIC ARM INTERACTION

Q	Questionnaire
Q1	How engaged did you feel during the collaboration with the robotic arm?
Q2	How well do you feel the robotic arm responded to your emotional expressions?
Q3	Did the responsiveness of the robotic arm enhance your overall engagement with the task?
Q4	How would you describe the speed of the robotic arm's movements during the task?
Q5	How would you describe the impact of your emotional expression on the speed of the robotic arm's movements?
Q6	How comfortable did you feel expressing your emotions during the collaboration with the robotic arm?
Q7	How satisfied were you with the overall collaboration experience?
Q8	How often did the speed of the robotic arm change during the task based on your emotional expression?

superior accuracy, a hypothesis that merits further research.

D. User Experience for Affective Robotic Arm

To assess the efficacy of the affective robotic arm using EmoFormer, we conducted a qualitative evaluation involving six participants. Each participant engaged in a collaborative task with the robotic arm, and their feedback was collected through a series of questions aimed at understanding their experience. The questions presented to the participants are summarized in Table VI.

The speed of a robotic arm is adjusted based on human emotions, the responsiveness and speed of interaction are crucial components of an interactive system. Real-time interaction, which provides immediate feedback, is highly valued as it enhances the user's perception of immersion and responsiveness [46]. We design our system by incorporating behaviors of robots based on the expression method in human-robot interactions [70]. When a robot detects a happy face, it moves more dynamically. This makes the interaction more engaging and responsive. On the other hand, When it detects negative emotions in the user's face, it slows down. This reflects cautious or hesitant movements. These behaviors can be programmed to simulate realistic emotional responses from the affective robotic arm. Specific design measures ensure the safe operation of our emotion-driven robotic arm. If no facial detection occurs, the system defaults to 'non-operation' mode for safety and improved user experience.

We evaluated the feedback from the participants. As shown in Table VII, almost all noticed frequent changes in the arm's speed based on their emotional cues. Every participant reported feeling highly engaged during their interaction. Participants found the robotic arm responsive to their emotional expressions, with many observing that the robot's responsiveness greatly enhanced their overall engagement with the task. The majority of participants felt there was a strong correlation between their emotional state and the speed of the robotic arm's movements. However, perceptions of the arm's speed varied. Some found it just right, while others felt it was too fast. Participants were comfortable expressing their emotions while interacting with the robot. The unanimous satisfaction with the overall collaboration experience highlights the potential advantages of the affective robotic arm.

TABLE IX
SCALE AND RESPONSES ON ROBOTIC ARM INTERACTION

Scale	1 or 2	3	4 or 5
Q1: 1 = Not Engaged, 5 = Highly Engaged	0	0	6
Q2: 1 = Poor, 5 = Excellent	0	2	4
Q3: 1 = No, not at all, 5 = Yes, Significantly	0	0	6
Q4: 1 = Very Slow, 5 = Very Fast	0	4	2
Q5: 1 = No correlation, 5 = Strongly correlated	0	1	5
Q6: 1 = Uncomfortable, 5 = Very Comfortable	0	1	5
Q7: 1 = Very Dissatisfied, 5 = Very Satisfied	0	0	6
Q8: 1 = Never, 5 = Almost All time	0	2	4

EmoFormer is used for affective Human-Robot Interaction. We evaluated the emotion-aware robotic arm through user experience. This showcases the potential of EmoFormer to enhance human-robot interactions. Furthermore, it opens the door to a multitude of future applications where emotion awareness can be harnessed to create more empathetic and affective robotic systems.

VI. CONCLUSION

In this research, we introduced the EmoFormer model for facial expression recognition, effectively addressing challenges related to accuracy and efficiency by leveraging Mix Transformer blocks and a novel fusion block. This method enables the model to create powerful representations and establish efficient connections among various facial regions, ensuring high accuracy in emotion classification. Our evaluation of EmoFormer on four datasets—FER2013, AffectNet-7, AffectNet-8, and DEAP—demonstrates its state-of-the-art performance. Specifically, our ensemble method achieved accuracies of 77.35% on FER2013, 67.71% on AffectNet-7, and 65.14% on AffectNet-8. For the DEAP dataset, the method achieved 98.07% accuracy for arousal and 97.86% for valence. A key application of our method involved implementing EmoFormer in an affective robotic arm. This implementation allows the system to adjust its speed based on the user's facial expressions. We validated this approach through a user experiment with six subjects. The results demonstrate the feasibility and effectiveness of our approach in creating emotionally intelligent human-robot interactions. Our system is capable of utilizing FER and integrating other modalities. Our future work aims to incorporate neurophysiological signals for multimodal emotion recognition, further advancing multimodal affective computing and significantly enhancing the field of affective HRI.

REFERENCES

[1] P. Ekman and H. Oster, "Facial expressions of emotion," *Annual Review of Psychology*, vol. 30, no. 1, pp. 527–554, Jan 1979.
[2] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *Journal of Research in Personality*, vol. 11, no. 3, pp. 273–294, Sep 1977.
[3] S. Zhang, Y. Zhang, Y. Zhang, Y. Wang, and Z. Song, "A dual-direction attention mixed feature network for facial expression recognition," *Electronics*, vol. 12, p. 3595, 2023.
[4] B. Wu and C. Lin, "Adaptive feature mapping for customizing deep learning based facial expression recognition model," *IEEE Access*, vol. 6, pp. 12 451–12 461, 2018.

[5] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proceedings of the International Conference on Learning Representations (ICLR)*, San Diego, CA, USA, May 7–9 2015, pp. 1–14, google Scholar. [Online]. Available: https://arxiv.org/abs/1409.1556

[6] 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 (CVPR)*, Las Vegas, NV, USA, June 27–30 2016, pp. 770–778, google Scholar. [Online]. Available: https://arxiv.org/abs/1512.03385

[7] L. Pham, T. H. Vu, and T. A. Tran, "Facial expression recognition using residual masking network," in *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 4513–4519.

[8] S. Vignesh, M. Savithadevi, M. Sridevi *et al.*, "A novel facial emotion recognition model using segmentation vgg-19 architecture," *International Journal of Information Technology*, vol. 15, pp. 1777–1787, 2023. [Online]. Available: https://doi.org/10.1007/s41870-023-01184-z

[9] C. Zheng, M. Mendieta, and C. Chen, "Poster: A pyramid cross-fusion transformer network for facial expression recognition," *arXiv preprint*, vol. arXiv:2204.04083, 2022. [Online]. Available: https://arxiv.org/abs/2204.04083

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[11] F. Ma, B. Sun, and S. Li, "Facial expression recognition with visual transformers and attentional selective fusion," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1236–1248, 2023.

[12] G. Pons and D. Masip, "Supervised committee of convolutional neural networks in automated facial expression analysis," *IEEE Transactions on Affective Computing*, vol. 9, no. 3, pp. 343–350, 2017.

[13] I. Goodfellow, D. Erhan, P. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, Y. Zhou, C. Ramaiah, F. Feng, R. Li, X. Wang, D. Athanasakis, J. Shawe-Taylor, M. Milakov, J. Park, and Y. Bengio, "Challenges in representation learning: A report on three machine learning contests," *Neural Networks*, vol. 64, 07 2013.

[14] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 12 077–12 090.

[15] S. Koelstra, C. Muehl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2012.

[16] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, jan 2019.

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

[18] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1. Ieee, 2005, pp. 886–893.

[19] I. Buciu and I. Pitas, "Application of non-negative and local non negative matrix factorization to facial expression recognition," in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, vol. 1. IEEE, 2004, pp. 288–291.

[20] S. H. Lee, K. N. Plataniotis, and Y. M. Ro, "Intra-class variation reduction using training expression images for sparse representation based facial expression recognition," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 340–351, 2014.

[21] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[22] Y. Tang, "Deep learning using linear support vector machines," *arXiv preprint arXiv:1306.0239*, 2013.

[23] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion aware facial expression recognition using cnn with attention mechanism," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2439–2450, 2018.

[24] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, "Region attention networks for pose and occlusion robust facial expression recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 4057–4069, 2020.

[25] F. Safavi and M. Rahnemoonfar, "Comparative study of real-time semantic segmentation networks in aerial images during flooding events," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 4–20, 2023.

[26] F. Safavi, T. Chowdhury, and M. Rahnemoonfar, "Comparative study between real-time and non-real-time segmentation models on flooding events," in *2021 IEEE International Conference on Big Data (Big Data)*, 2021, pp. 4199–4207.

[27] F. Safavi, K. Patel, and R. K. Vinjamuri, "Towards efficient deep learning models for facial expression recognition using transformers," in *2023 IEEE 19th International Conference on Body Sensor Networks (BSN)*, 2023, pp. 1–4.

[28] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1925–1934.

[29] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[30] K.-H. Pong and K.-M. Lam, "Multi-resolution feature fusion for face recognition," *Pattern Recognition*, vol. 47, no. 2, pp. 556–567, 2014.

[31] J. Shao and Y. Qian, "Three convolutional neural network models for facial expression recognition in the wild," *Neurocomputing*, vol. 355, pp. 82–92, 2019.

[32] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part VII*. Berlin, Heidelberg: Springer-Verlag, 2018, p. 3–19.

[33] C. Hewitt and H. Gunes, "Cnn-based facial affect analysis on mobile devices," *arXiv preprint arXiv:1807.08775*, 2018.

[34] P. Barros, N. Churamani, and A. Sciutti, "The facechannel: a fast and furious deep neural network for facial expression recognition," *SN Computer Science*, vol. 1, no. 6, p. 321, 2020.

[35] 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*, vol. 35, no. 4, 2021, pp. 3510–3519.

[36] J. Mao, R. Xu, X. Yin, Y. Chang, B. Nie, and A. Huang, "Poster++: A simpler and stronger facial expression recognition network," *arXiv preprint arXiv:2301.12149*, 2023.

[37] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *ArXiv*, vol. abs/2010.11929, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:225039882

[38] F. Xue, Q. Wang, and G. Guo, "Transfer: Learning relation-aware facial expression representations with transformers," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 3581–3590.

[39] Z. Wen, W. Lin, T. Wang, and G. Xu, "Distract your attention: Multi-head cross attention network for facial expression recognition," *Biomimetics*, vol. 8, no. 2, p. 199, 2023.

[40] F. Safavi, P. Olikkal, D. Pei, S. Kamal, H. Meyerson, V. Penumalee, and R. Vinjamuri, "Emerging frontiers in human–robot interaction," *Journal of Intelligent & Robotic Systems*, vol. 110, no. 2, p. 45, 2024.

[41] Q. Wei, T. Li, and D. Liu, "Learning control for air conditioning systems via human expressions," *IEEE Transactions on Industrial Electronics*, vol. PP, pp. 1–1, 06 2020.

[42] N. Zhao, D. Lu, K. Hou, M. Chen, X. Wei, X. Zhang, and B. Hu, "Fatigue detection with spatial-temporal fusion method on covariance manifolds of electroencephalography," *Entropy*, vol. 23, no. 10, 2021. [Online]. Available: https://www.mdpi.com/1099-4300/23/10/1298

[43] H. Abdollahi, M. Mahoor, R. Zandie, J. Sewierski, and S. Qualls, "Artificial emotional intelligence in socially assistive robots for older adults: A pilot study," *IEEE Transactions on Affective Computing*, pp. 1–1, 2022.

[44] 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 Transactions on Affective Computing*, vol. 13, no. 4, pp. 2132–2143, 2022.

[45] Y. Zhang, M. Z. Hossain, and S. Rahman, "Deepvanet: A deep end-to-end network for multi-modal emotion recognition," in *Human-Computer*

Interaction – INTERACT 2021: 18th IFIP TC 13 International Conference, Bari, Italy, August 30 – September 3, 2021, Proceedings, Part III, 2021, pp. 227–237.

[46] J. Steuer, F. Biocca, M. R. Levy *et al.*, "Defining virtual reality: Dimensions determining telepresence," *Communication in the age of virtual reality*, vol. 33, no. 37-39, p. 1, 1995.

[47] R. Lavoie, K. Main, C. King, and D. King, "Virtual experience, real consequences: the potential negative emotional consequences of virtual reality gameplay," *Virtual Reality*, vol. 25, no. 1, pp. 69–81, 2021.

[48] 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. 26 756–26 768, 2022.

[49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[50] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2261–2269.

[51] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual Attention Network for Image Classification," *CoRR*, vol. abs/1704.0, 2017. [Online]. Available: http://arxiv.org/abs/1704.06904

[52] S. Woo, J. Park, J.-Y. Lee, and I.-S. Kweon, "CBAM: Convolutional Block Attention Module," in *European Conference on Computer Vision*, 2018.

[53] R. Pecoraro, V. Basile, and V. Bono, "Local multi-head channel self-attention for facial expression recognition," *Information*, vol. 13, no. 9, p. 419, 2022.

[54] A. Howard, R. Pang, H. Adam, Q. Le, M. Sandler, B. Chen, W. Wang, L.-C. Chen, M. Tan, G. Chu, V. Vasudevan, and Y. Zhu, "Searching for MobileNetV3," 2019, pp. 1314–1324.

[55] J. Yang, Z. Lv, K. Kuang, S. Yang, L. Xiao, and Q. Tang, "RASN: Using Attention and Sharing Affinity Features to Address Sample Imbalance in Facial Expression Recognition," *IEEE Access*, vol. 10, pp. 103 264–103 274, 2022.

[56] X. Liang, J. Liang, T. Yin, and X. Tang, "A lightweight method for face expression recognition based on improved MobileNetV3," *IET Image Processing*, vol. 17, no. 8, pp. 2375–2384, 2023. [Online]. Available: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/ipr2.12798

[57] J. Yang, Z. Lv, K. Kuang, S. Yang, L. Xiao, and Q. Tang, "RASN: Using Attention and Sharing Affinity Features to Address Sample Imbalance in Facial Expression Recognition," *IEEE Access*, vol. 10, pp. 103 264–103 274, 2022.

[58] S. Zhang, Y. Zhang, Y. Zhang, Y. Wang, and Z. Song, "A dual-direction attention mixed feature network for facial expression recognition," *Electronics*, vol. 12, no. 17, p. 3595, 2023.

[59] Z. Wen, W. Lin, T. Wang, and G. Xu, "Distract your attention: Multi-head cross attention network for facial expression recognition," *Biomimetics*, vol. 8, no. 2, p. 199, 2023.

[60] 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.

[61] 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*, vol. 35, no. 4, 2021, pp. 3510–3519.

[62] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[63] Y. An, N. Xu, and Z. Qu, "Leveraging spatial-temporal convolutional features for eeg-based emotion recognition," *Biomedical Signal Processing and Control*, vol. 69, p. 102743, 2021.

[64] M. A. Ozdemir, M. Degirmenci, E. Izci, and A. Akan, "Eeg-based emotion recognition with deep convolutional neural networks," *Biomedical Engineering/Biomedizinische Technik*, vol. 66, no. 1, pp. 43–57, 2021.

[65] V. Padhmashree and A. Bhattacharyya, "Human emotion recognition based on time–frequency analysis of multivariate eeg signal," *Knowledge-Based Systems*, vol. 238, p. 107867, 2022.

[66] E. S. Salama, R. A. El-Khoribi, M. E. Shoman, and M. A. W. Shalaby, "A 3d-convolutional neural network framework with ensemble learning techniques for multi-modal emotion recognition," *Egyptian Informatics Journal*, vol. 22, no. 2, pp. 167–176, 2021.

[67] S. Wang, J. Qu, Y. Zhang, and Y. Zhang, "Multimodal emotion recognition from eeg signals and facial expressions," *IEEE Access*, vol. 11, pp. 33 061–33 068, 2023.

[68] Y. Wu and J. Li, "Multi-modal emotion identification fusing facial expression and eeg," *Multimedia Tools and Applications*, vol. 82, no. 7, pp. 10 901–10 919, 2023.

[69] P. Ekman and E. L. Rosenberg, *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997.

[70] L. Li and Z. Zhao, "Designing behaviors of robots based on the artificial emotion expression method in human–robot interactions," *Machines*, vol. 11, no. 5, p. 533, 2023.