IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 1 October 2023, accepted 15 October 2023, date of publication 18 October 2023, date of current version 3 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3325407

RESEARCH ARTICLE

Enhancing Facial Expression Recognition System in Online Learning Context Using Efficient Deep Learning Model

MOHAMMED ALY[1], **ABDULLATIF GHALLAB**[2], **AND ISLAM S. FATHI**[3,4]

[1]Department of Artificial Intelligence, Faculty of Artificial Intelligence, Egyptian Russian University, Badr City 11829, Egypt
[2]Department of Computer Science, Faculty of Computing and Information Technology, University of Science and Technology, Sana'a, Yemen
[3]Department of Computer science and AI applications, Faculty of Computers and Information, El Saleheya El Gadida University, El Saleheya El Gadida 44813, Egypt
[4]Department of Information Systems, Al Alson Higher Institute, Cairo 11762, Egypt

Corresponding author: Abdullatif Ghallab (ghallababdullatif@gmail.com)

ABSTRACT This paper presents an online educational platform that leverages facial expression recognition technology to monitor students' progress within the classroom. Periodically, a camera captures images of students in the classroom, processes these images, and extracts facial data through detection methods. Subsequently, students' learning statuses are assessed using expression recognition techniques. The developed approach then dynamically refines and enhances teaching strategies using the acquired learning status data. In the course of the experiment, we enhance facial expression recognition accuracy through the utilization of ResNet-50 for effective feature extraction. Additionally, by adjusting the residual down-sampling module, we bolster the correlation among input features, thus mitigating the loss of feature information. Simultaneously, a convolutional attention mechanism module is incorporated to reduce the influence of irrelevant areas within the feature map. The proposed method achieves an accuracy of 87.62% and 88.13 % on the RAF-DB and FER2013 expression datasets, respectively. In comparison with the original ResNet-50 network and the expression recognition outcomes found in existing literature, the suggested approach demonstrates enhanced accuracy and improved detection of students' learning states and expression variations. Consequently, the application of facial expression recognition technology in online learning, along with the optimization of online teaching resources and strategies grounded in the results of recognition, holds tangible value for augmenting the quality of online learning experiences. We have benchmarked the proposed model against state-of-the-art techniques and conducted evaluations using the FER-2013, CK+, and KDEF datasets. The significance of these results lies in their potential application within educational institutions.

INDEX TERMS Facial expression recognition, online learning, ResNet, residual network, attention mechanism.

I. INTRODUCTION

Amidst the swift evolution and unceasing enhancement of artificial intelligence, online learning has entrenched itself as the standard across diverse educational realms. In contrast to the conventional in-person classroom educational model, online instruction transcends temporal and geographical constraints, empowering students to cultivate autonomous learning. Furthermore, online educational platforms boast a profusion of instructional resources, amplifying students' learning efficacy and kindling their intellectual curiosity [1].

Nonetheless, while the prevalence of online classroom platforms persists, several constraints emerge: Firstly, digital education sidelines face-to-face instruction, rendering it incapable of furnishing immediate succor to students grappling with comprehension hurdles, thereby fostering negative sentiments and exhaustion. Secondly, within online courses, the detachment of students' emotions and actions from the learning environment often precludes instructors from

The associate editor coordinating the review of this manuscript and approving it for publication was Joewono Widjaja.

administering timely guidance and rectification, ultimately undermining the potency of students' learning endeavors. Thirdly, the majority of online classroom platforms gauge students' scholastic achievements via post-lesson exercises, a mode manifestly ill-suited to contemporary practical requisites and inadequate in delivering a comprehensive and unbiased account of students' educational accomplishments [2].

Hence, the acquisition of students' learning behaviors within online classrooms, their subsequent analysis to ascertain classroom engagement, and the establishment of a robust assessment mechanism stand as pivotal cornerstones in realizing scientifically-driven online education and heightening pedagogical effectiveness [3], [4].

In recent times, spurred by advancements in facial expression recognition technology, researchers globally have harnessed this technology to monitor students' classroom engagement. Utilizing students' facial expression data as indicators, they strive to discern the dynamics of their classroom involvement [5], [6]. For instance, in [7], an emotion recognition module tailored for distance learners was devised. This module, driven by intelligent agents, employs expression recognition and video tracking technology to enhance the accuracy of gauging learners' emotional states within remote learning environments. In [8], authors integrated expression recognition technology into the teacher-student emotional interaction subsystem of the three-dimensional virtual learning platform Magic Learning. This integration facilitated emotional recognition and intervention capabilities, amplifying the platform's efficacy for learners. Furthermore, author successfully amalgamated the three-dimensional learning state space with emotional dimensions, culminating in an effective method for categorizing classroom states. Nevertheless, owing to the intricate backdrop of facial features and the current limitations of algorithms, the application of facial expression recognition in intelligent education remains a work in progress. The actual impact of models designed to recognize and analyze students' facial expressions within the classroom context falls considerably short of the desired ideal. The primary challenges encompass:

1) The facial expression recognition technology employed in monitoring online classrooms operates within its own unique network; however, it grapples with the drawbacks of employing large models and yielding limited recognition accuracy. Enhancing the precision of learning-based emotional facial expression recognition algorithms remains an unresolved issue necessitating further investigation. Central to the quandary of learning emotional facial expression recognition lies the extraction of facial expression features. Devising methods to proficiently extract crucial facial expression attributes, curbing noise and superfluous data during feature extraction, thereby ameliorating recognition accuracy, all hinges on pioneering innovations within the realm of facial expression recognition algorithms.
2) For personalized online learning, a proficient system for tracking students' learning progress and delivering intelligent feedback and guidance is conspicuously absent. Presently, the majority of learning progress tracking systems operate from the teacher's vantage point, wherein instructional adjustments are orchestrated based on the monitoring outcomes furnished by the system.

In order to address these issues, this study formulates an online classroom monitoring system grounded in an enhanced expression recognition technique. This approach involves refining the ResNet deep convolutional network, fine-tuning its residual modules, and integrating the Convolutional block attention mechanism (CBAM) attention mechanism to enhance the precision of expression recognition. By scrutinizing students' emotional states, appraising classroom dynamics, and furnishing relevant cues, this system pioneers a novel methodology for assessing students' engagement within the online classroom setting. It establishes a framework that aids educators in overseeing students' progress during their online learning journey. Consequently, this research domain assumes significant research and practical significance.

II. RELATED WORK

A. RESEARCH OF FACIAL EXPRESSION RECOGNITION

Since 2006, the surge of interest in deep learning has been driven by its capacity to adeptly extract nonlinear features. Researchers have pioneered various enduring network architectures using deep learning, including renowned ones like AlexNet, VGGNet, and GoogLeNet. In [9], an enhanced approach involving multi-channel convolution, global average pooling, and batch normalization was incorporated into the AlexNet network for facial expression recognition, leading to a remarkable 13.24% boost in accuracy compared to the original network configuration. Similarly in [10], authors employed an expression recognition methodology rooted in VGGNet, achieving an enhanced expression recognition rate. In contrast to conventional techniques, deep learning-driven approaches consistently yield superior detection accuracy [11]. However, this heightened accuracy comes hand in hand with prolonged processing times, as the utilization of extensive convolutional networks housing numerous parameters inevitably elongates processing times and escalates network intricacy, rendering the training of such deep convolutional neural networks a formidable task. Against this backdrop, He Kaiming et al. [12] introduced the ResNet network model, distinguished by its novel integration of residual concepts within the convolutional neural network framework. The incorporation of Residual Units indeed empowers the training of convolutional neural networks with depths reaching up to 152 layers. In totality, ResNet excels in achieving low error rates, demanding fewer parameters, imposing diminished computational complexity, and expediting model training—an outstanding solution for training deep learning models [13].

In conventional classrooms, teachers and students engage in face-to-face communication within the same physical space. However, the landscape of online educational platforms necessitates feedback mechanisms rooted in facial expression recognition. This realm has garnered extensive attention and scrutiny from numerous experts and scholars.

Mahmoud Neji's online learning platform, founded on the tenets of affective computing, adeptly discerns six fundamental emotions: happiness, disgust, anger, surprise, sadness, and fear [14].

Kiavash Bahrini [15], on the other hand, melded facial expression recognition with speech analysis to introduce the FILTWAM learning framework. Throughout the learning journey, learners' vocal cues and facial expressions are captured through devices like microphones and cameras.

Zhan Zehui innovated an emotional cognitive recognition model tailored for remote learners, incorporating emotion recognition technology alongside eye-tracking mechanisms [7].

Delving into the terrain of individual emotional needs within online learning, Wang Zhiliang and team delved into an intricate web of ideas, system models, and technical strategies, culminating in a harmonious interaction paradigm connecting learning platforms and learners [13].

In Ning Zhou et al.'s study [16], the primary focus was on multitask learning to enhance classification results. In this paper, they introduced and developed a lightweight Convolutional Neural Network (CNN) to achieve real-time emotion recognition and aggregate facial emotion detection. Their approach involves performing face identification using Multi-Task Cascaded Convolutional Neural Networks (MTCNN). Subsequently, the resulting facial coordinates are fed into the initially created facial emotion classification model, enabling emotion classification. One noteworthy aspect is that this model leverages certain features from the cascade detection in multi-task cascaded convolutional networks, allowing for reduced memory usage. In our expression categorization model, the conventional fully connected layer is replaced by Global Average Pooling, resulting in a more transparent relationship between feature map channels and corresponding categories. Additionally, this model incorporates depth-wise separable convolutions and residual modules, leading to a significant reduction in the number of parameters and increased portability. As a result, it yields strong recognition performance even when faced with images beyond the provided datasets.

Eric Granger et al.'s study [17] primarily focused on recognizing facial expressions linked to depressive behavior. This paper presents an innovative approach. Unlike conventional models that employ 2D convolutional neural networks, this study employs a 3D convolutional structure to capture information related to the severity of depression accurately. The distinctive model employed in this research, known as the Multiscale Spatiotemporal Network (MSN), adeptly analyzes depressive emotions from both images and videos. To evaluate its performance, the model is rigorously assessed and compared to other state-of-the-art models such as C3D, utilizing the AVEC2013 and AVEC2014 datasets. The experimental results demonstrate a remarkably high level of accuracy.

In Liqian Liang et al.'s study [18], their primary focus was on recognizing not only the six basic expressions but also various sub-emotions within each basic emotion category. Unlike most other papers that typically address only the fundamental emotions, this research delves comprehensively into the diverse spectrum of human emotions by identifying distinct sub-emotions within each basic category. The paper places significant emphasis on emotion recognition from both images and videos, introducing two novel concepts. Firstly, they introduce their unique dataset, the Fine-grained Emotion (FG) dataset, which is meticulously curated under unconstrained conditions. Secondly, they introduce two innovative neural networks: the Multi-Scale Action Unit Network (MSAU-Net) for image-based emotion recognition and the Two-stream Multi-Scale Action Unit Network (TMSAU-Net) for video-based emotion recognition. In TMSAU-Net, they integrate both an attention mechanism and a temporal stream branch, facilitating the simultaneous learning of spatial and temporal features. Moreover, for fine-grained emotion analysis, they incorporate progressive learning techniques. When compared to benchmark datasets, their experimentation results demonstrate the superior performance of the proposed method, particularly when evaluated against the FG-Emotion dataset.

In the research conducted by Y. ELsayed et al. [19], the authors introduced a novel hybrid convolutional neural network, augmented by a local binary pattern, for the precise extraction of features, with a particular focus on effectively recognizing emotions in individuals wearing face masks. The study aims to classify the seven fundamental emotions, including anger, happiness, sadness, surprise, contempt, disgust, and fear. The proposed method underwent testing on two distinct datasets: the first dataset encompasses CK and CK+, while the second dataset comprises M-LFW-FER. The results obtained from these experiments reveal that the emotion recognition, even in cases where individuals are wearing face masks, achieved an impressive accuracy rate of 70.76% across three different emotions.

In Kunyoung Lee et al.'s study [20], the authors presented a novel approach aimed at optimizing the combined learning of discrete and continuous emotional states within eight distinct expression classes, encompassing valences and arousal levels. Their proposed knowledge distillation model leverages Emonet, a cutting-edge continuous estimation method, as the teacher model, while employing a lightweight network as the student model. Interestingly, the study demonstrated that the performance degradation can be kept to a minimum, despite the student models involving multiply-accumulate operations of approximately 3.9 G and 0.3 G when employing EfficientFormer and MobileNetV2, respectively. This is significantly less computational load compared to the teacher model, which requires 16.99 G of computation. Additionally, this approach led to substantial enhancements in computational efficiency, achieving a 4.35-fold and 56.63-fold increase in efficiency using EfficientFormer and MobileNetV2, respectively. Furthermore, the reduction in facial expression classification accuracy was approximately 1.35% and 1.64%, respectively, showcasing the effectiveness of this method.

In the research conducted by Jiarui Zhong et al. [21], the authors employed a Northern Goshawk optimization (NGO) algorithm to fine-tune the hyperparameters of the BILSTM network, specifically for the task of facial expression recognition. Their proposed methodologies underwent thorough evaluation and comparison with other existing approaches across multiple datasets, including FER2013, FERplus, and RAF-DB, while considering factors such as cultural background, race, and gender. The outcomes of their study indicate that the recognition accuracy achieved by their model on the FER2013 and FERPlus datasets significantly surpasses that of the conventional VGG16 network, highlighting the effectiveness of their optimized approach.

In the research conducted by Andrada-Livia Cîrneanu et al. [22], this paper undertook a comprehensive review of the most recent advancements in the field of Facial Expression Recognition (FER). The primary focus was on contemporary neural network models that implement specialized facial image analysis algorithms for the detection and identification of facial emotions. The overarching objective of this paper is to provide an in-depth exploration of the historical and conceptual evolution of neural network architectures that have demonstrated significant accomplishments in the realm of FER. This paper advocates for the adoption of convolutional neural network (CNN)-based architectures, emphasizing their superiority over alternative neural network designs like recurrent neural networks or generative adversarial networks. It meticulously delineates the essential components and performance attributes of each architectural approach, as well as elucidating the strengths and limitations of the proposed models found in the reviewed literature. Moreover, this paper meticulously catalogs the existing datasets currently employed for the recognition of emotions conveyed through facial expressions, including micro-expressions. Furthermore, it underscores the pervasive utilization of FER systems across diverse domains such as healthcare, education, security, and social IoT, highlighting their multifaceted applicability.

In the research conducted by Andrada-Livia Cîrneanu et al. [23], the authors introduced an enhanced version of the YOLOv5 network by incorporating attention mechanisms into the Backbone model of YOLOv5. In their experiments, they systematically explored the impact of various attention mechanisms on YOLOv5. These attention mechanisms were added after each CBS module within the CSP1_X structure of the Backbone section. Additionally, attention mechanisms were integrated into different segments of the Focus, CBS, and SPP modules of YOLOv5 to assess their effects on these individual modules. The results of their investigation revealed that the network incorporating coordinate attentions after each CBS module in the CSP1_X structure achieved a detection time of 25 ms and an accuracy of 77.1%, marking a notable improvement of 3.5% compared to the standard YOLOv5. It outperformed several other networks, including Faster-RCNN, R-FCN, ResNext-101, DETR, Swin-Transformer, YOLOv3, and YOLOX. Moreover, as an application of their enhanced model, the authors designed a real-time facial expression recognition system for teachers. This system employs a camera and teaching video to detect and analyze the temporal distribution of teachers' facial expressions, providing valuable insights in real-time.

III. MATERIALS AND METHODS

Facial Emotion Recognition (FER) has emerged as a domain of interdisciplinary research. Beyond its various applications, FER holds significant relevance in the realm of security, where it can be employed to identify and authenticate individuals' emotions in photographs or videos.

A. DATASETS

In order to assess and evaluate various methods for classifying and recognizing facial emotions, the use of standardized datasets was essential. Over the past few decades, several facial emotion datasets have been enhanced. The subsequent sections will provide a comprehensive overview of some of the standardized benchmark datasets utilized in this study.

1) FER-2013

The FER-2013 dataset comprises 33,000 grayscale facial images depicting seven fundamental emotions: neutrality, happiness, anger, sadness, surprise, disgust, and fear [24]. These images have been automatically aligned to ensure that each face is roughly centered and occupies a similar amount of space within the frame.

2) 2 CK+

CK+ comprises 593 images depicting 120 individuals aged between 18 and 30 years [25]. This dataset encompasses images representing all seven primary emotions, with resolutions of 640 × 490 or 640 × 480 pixels in 8-bit grayscale. The dataset's demographic distribution consists of approximately 81% European-Americans, 13% African-Americans, and 6% from various other ethnic backgrounds, maintaining a gender ratio of 65% women to 35% men.

3) THE KDEF

The KDEF dataset [26] comprises 490 JPEG images, featuring 35 women and 35 men, each displaying seven distinct emotional expressions at a resolution of 72 × 72 pixels.

4) RAF-DB

RAF-DB represents a real-world facial expression dataset gathered from online sources, consisting of a grand total of 29,672 facial images. These images are meticulously annotated with seven fundamental emotional categories as well as eleven intricate compound emotional labels. Within the RAF-DB dataset, there are 12,271 samples designated for training purposes, distributed across emotions including surprise (1,290 samples), fear (281 samples), disgust (717 samples), happiness (4,772 samples), sadness (1,982 samples), anger (705 samples), and neutrality (2,524 samples). Additionally, there are 3,068 testing samples provided for the purpose of evaluating Facial Expression Recognition (FER) systems.

RAF-DB stands out due to its remarkable diversity in real-world scenarios and challenging conditions, making it an invaluable resource for assessing the effectiveness of FER techniques under various factors such as differing poses, lighting conditions, and occlusions [27].

IV. METHODOLOGY

Within this segment, the authors initially present enhancements made to the ResNet network model. These augmentations encompass the integration of the CBAM attention mechanism and the refinement of the network's residual modules, collectively aimed at elevating expression recognition accuracy. Subsequently, the authors outline the architecture of the online learning status monitoring system predicated on this refined expression recognition methodology. Lastly, the authors expound upon the assessment of online learning status.

A. IMPROVING THE NETWORK STRUCTURE OF EXPRESSION RECOGNITION

1) CONVOLUTIONAL BLOCK ATTENTION MECHANISM (CBAM)

The fundamental concept of the attention mechanism lies in assigning weights that empower neural networks to concentrate on pertinent information, downplay less crucial details, and alleviate data noise. This paper seeks to enhance the profundity of deep network models by incorporating attention mechanisms to emphasize significant attributes while mitigating the significance of extraneous elements. Attention mechanisms can be classified into Spatial Attention Mechanism (SAM) [28], Channel Attention Mechanism (CAM) [29], and a Hybrid Attention Mechanism. CAM is often manifested as weights, where the magnitude of weight directly corresponds to the level of attention allocated to the associated region. Nonetheless, in image recognition tasks, grappling with the ramifications of rotation, distortion, and scale variations poses a challenge. Here, SAM proves effective in preserving pivotal image information and mitigating the impact of operations like transformations.

CBAM amalgamates CAM and SAM within a sequential framework, yielding superior outcomes compared to models reliant solely on CAM. The configuration of the CBAM network is depicted in Fig. 1, where F signifies the feature image derived from the convolution layer, Mc(F) represents the generated channel attention image, F' corresponds to the feature image obtained by multiplying F with Mc(F), Ms(F') signifies the spatial attention image generated, and F'' denotes the feature image derived from the multiplication of F' and Ms(F').

The CBAM, residual module adjustment, and the importance of bolstering correlation among input features in the context of using ResNet-50 for effective feature extraction in facial expression recognition. As shown in Fig.1. The CBAM network structure comprises the following components:

Channel-wise Attention: CBAM's channel-wise attention mechanism focuses on feature channels within each feature map. It calculates the importance of each channel by aggregating information globally across spatial dimensions. This helps the network to dynamically weigh the relevance of each feature channel for a given task. In facial expression recognition, this means paying more attention to channels that encode important facial features (e.g., eyes, nose, mouth) while downweighting irrelevant channels.

- **Spatial Attention:** CBAM's spatial attention mechanism, on the other hand, highlights relevant spatial regions within each feature map. It calculates the importance of each spatial position by considering both average-pooled and max-pooled features. This enables the network to attend to specific regions of the feature map where important facial expressions cues are localized.
- **Combined Attention**: By combining both channel-wise and spatial attention, CBAM creates a feature map that is more focused on relevant channels and spatial regions. This results in a more discriminative feature representation for facial expression recognition, as the network emphasizes both what features to pay attention to (channels) and where to pay attention (spatial regions).

2) RESIDUAL MODULE ADJUSTMENT

- **Residual Blocks in ResNet:** ResNet architectures use residual blocks to facilitate training of very deep neural networks. These blocks typically include two convolutional layers followed by an element-wise addition operation with a shortcut connection (skip connection). While residual blocks are effective, they often include downsampling layers (e.g., max-pooling or strided convolutions) that reduce the spatial resolution of feature maps.
- **Issue with Downsampling:** Downsampling can lead to a loss of fine-grained information, which is crucial in facial expression recognition. Subtle facial cues, such as the movement of facial muscles, are often spatially detailed and may be lost during downsampling. This can hinder the network's ability to recognize certain expressions accurately.

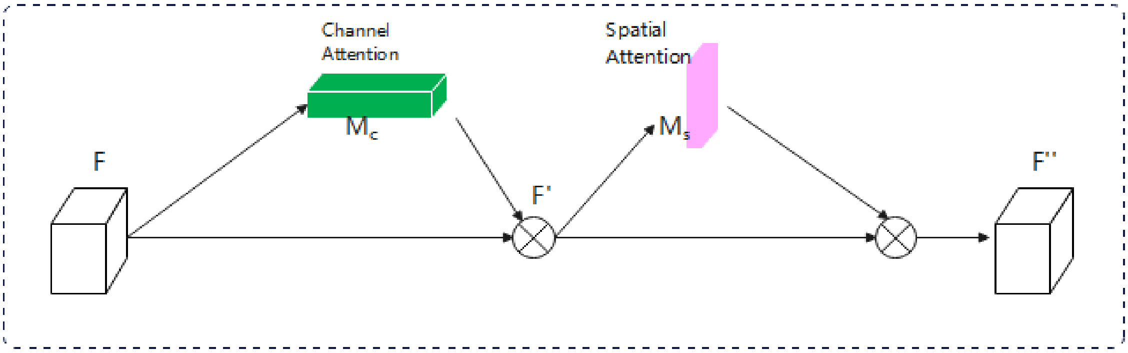


FIGURE 1. **CBAM network structure.**

- **Residual Down-sampling Module Adjustment:** To address this issue, you can adjust the residual down-sampling module by modifying the down-sampling layers. Instead of aggressive downsampling, you can use techniques such as:
- **Dilated Convolution:** Dilated (atrous) convolution can maintain spatial resolution while increasing the receptive field of the network. It allows the network to capture information from a broader context without significant downsampling.
- **Strided Convolutions with Padding:** You can use strided convolutions with appropriate padding to control the amount of downsampling. This helps in preserving spatial information.
- **Preserving Correlation and Feature Information:** These adjustments aim to bolster the correlation among input features. When facial features (e.g., eyes, nose, mouth) are closely correlated in the input image, these correlations are preserved as the features pass through the network. This is critical because facial expressions are characterized by the spatial relationships among these features. By mitigating the loss of feature information, the network can better capture these spatial dependencies.

In summary, CBAM enhances feature extraction by dynamically emphasizing relevant channels and spatial regions. Adjusting the residual down-sampling module in ResNet-50 mitigates the loss of fine-grained feature information and bolsters the correlation among input features, which is essential for accurate facial expression recognition. These modifications collectively improve the network's ability to capture and represent facial expression cues effectively.

Fig. 2 illustrates the ResNet-50 architecture, predominantly composed of the input, convolutional, and output components. Within this depiction, layers conv3 through conv5 encompass a down-sampled residual module. Specifically focusing on the conv3 layer, both the down-sampled residual module and the standard residual module are showcased in Fig. 3. While the residual modules for conv4 and conv5 bear resemblances, disparities emerge in terms of channel numbers.

The input data of the down-sampled residual module undergoes convolution with a 1 × 1 kernel at a stride of 2, leading to a halving of the feature image's height and width. This, however, results in the disregard of 3/4 of the feature information, consequently discarding vital details and exerting an impact on ensuing feature extraction. In response, this study refines the down-sampled residual module. Specifically, preceding the 1 × 1 convolution, a 2 × 2 average pooling layer with a stride of 2 is introduced, accompanied by a change in the original convolution layer's stride to 1. Simultaneously, a convolution attention module is integrated into the residual block of ResNet, augmenting the model's capability to extract pivotal features. This augmented structure is termed CBAM-Bottleneck. The modified residuals are depicted in Fig. 4.

In an endeavor to enhance the network's performance without incurring overfitting from excessive model parameters during training, adaptations are made to ResNet's down-sampled residual module, alongside the addition of a CBAM attention mechanism. The resulting network configuration is delineated in Fig. 5. Upon feeding the pre-processed image into the network model and transmitting it through the initial convolutional layer, facial expression characteristics are distilled via an advanced residual block operation. Incorporating residual blocks of varying dimensions imbues neural networks with increased depth, facilitating the extraction of more intricate and intricate attributes, while mitigating the challenge of gradient vanishing through inter-layer connections. The introduction of the CBAM attention module contributes to augmenting the channel and spatial dimensions of locally significant features, thereby expediting the model's convergence and elevating its recognition accuracy.

B. SYSTEM ARCHITECTURE

The online learning system constitutes an emotional analysis model that harmonizes online learning with expression recognition technology. It acquires students' facial images via a camera, applies facial expression recognition technology to categorize these expressions, and subsequently evaluates

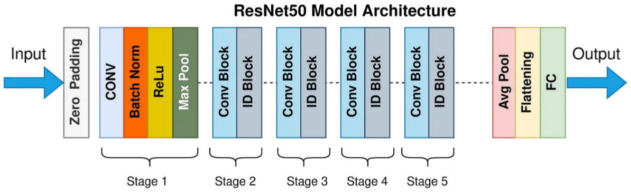


FIGURE 2. ResNet-50 network structure [30], [31].

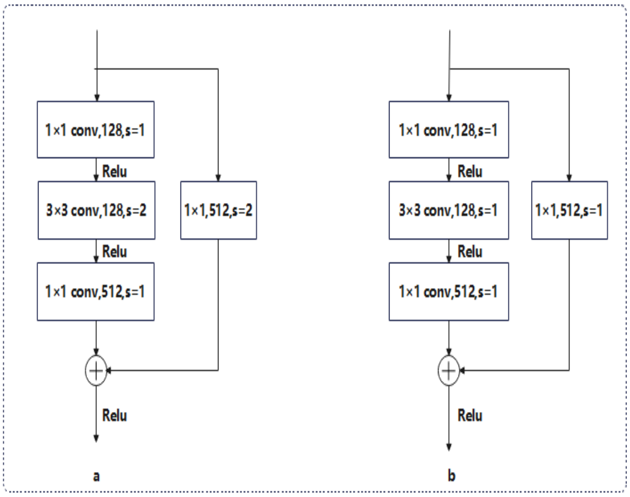


FIGURE 3. CBAM network structure [31].

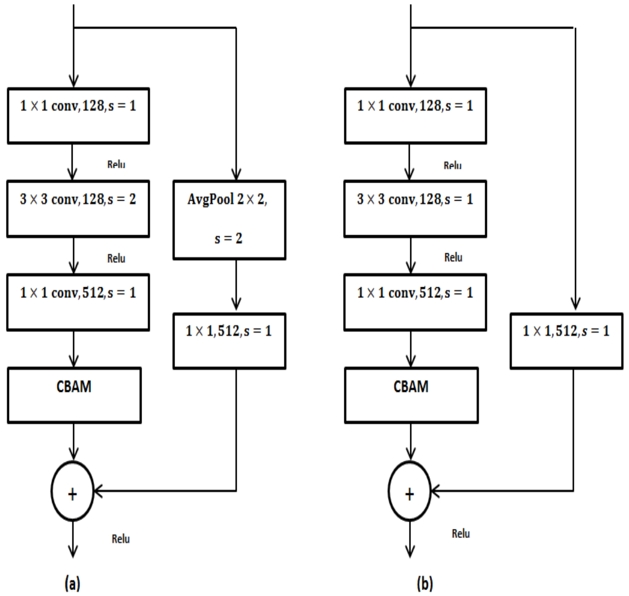


FIGURE 4. Improved sub-sampled and normal residual module.

students' learning states by considering the predefined "emotional state" classification strategy. The system then provides feedback to the students through the platform. The design process of the suggested online learning system is visualized in Fig. 6.

The entire system workflow is outlined as follows:

(1) User registration and login.

(2) Learner's learning phase. Learners engage with content recommended by the system or sourced through their own searches.

(3) Emotion recognition. During learners' educational journey, the system employs a camera to periodically capture facial images at preset intervals. It then proceeds to detect the learners' facial expressions and gauge their emotional state while learning. If their emotional state is positive, they proceed to step (2) for continuous learning. Otherwise, they advance to step (4) for instructional adjustments. (4) Adjustment of learning strategies. Learners' curriculum is recalibrated, potentially involving measures such as reducing difficulty, moderating learning pace, or interspersing extracurricular knowledge, such as humanistic contexts tied to specific knowledge points or engaging narratives, aimed at ameliorating learners' emotional disposition. The recognition of emotional states stands as a pivotal facet within the online learning system, with its structural layout portrayed in Fig. 7.

The crux of the emotion recognition module lies in the expression recognition component, its fundamental procedure depicted in Fig. 8. The sequential course of action is as follows: (1) Capture learners' facial data via the camera. (2) Apply image pre-processing. (3) Detect and localize the face within the image. If no face is detected, revert to step (1). Alternatively, advance to step (4). (4) Extract facial expression attributes and classify the expression. (5) Generate output indicating the outcome of expression recognition.

C. ONLINE LEARNING STATUS ASSESSMENT

For individuals engaged in online learning, the criteria governing the evaluation and assessment of their learning behaviors are outlined as follows:

1. **Focus:** During online learning, students' facial presence is monitored every 10 seconds. If their face is undetected for more than half of this duration, it is inferred that they are either inattentive or facing technical issues.
2. **Difficulty:** Emotions exhibited by students are analyzed throughout their online learning experience. If a student displays consecutive instances of anger or distress emotions (30 times in a row), and the system discerns that the corresponding knowledge points are challenging, the system will proactively provide a comprehensive interpretation of those points or arrange for immediate online teacher support.
3. **Engagement:** Emotions of students are scrutinized during their online learning sessions, and their engagement level is gauged based on the proportion of happy or surprised emotions detected.
4. **Involvement Rate:** By analyzing students' online learning behavior, the system evaluates whether students are active or passive learners based on the proportion of time spent with their heads bowed. If students consistently bow their heads for more than 50%

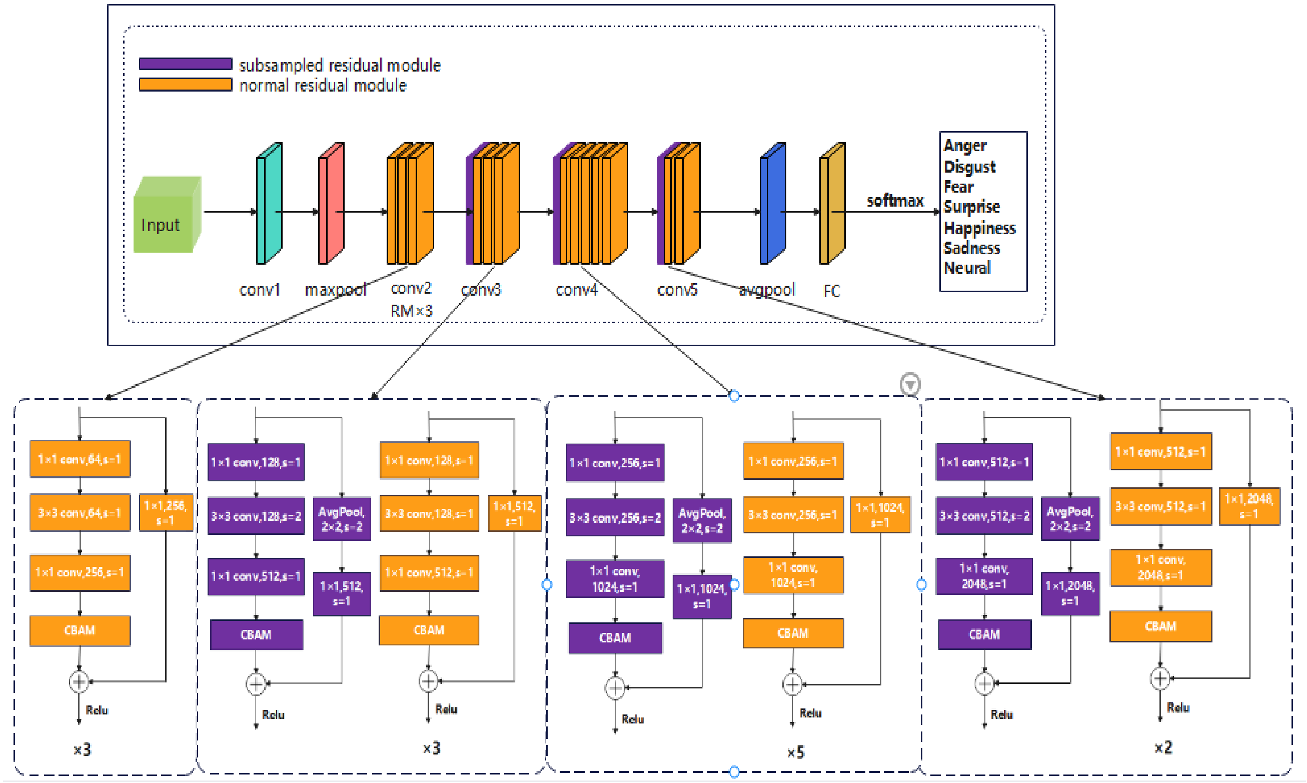


FIGURE 5. ResNet50 network embedded in CBAM.

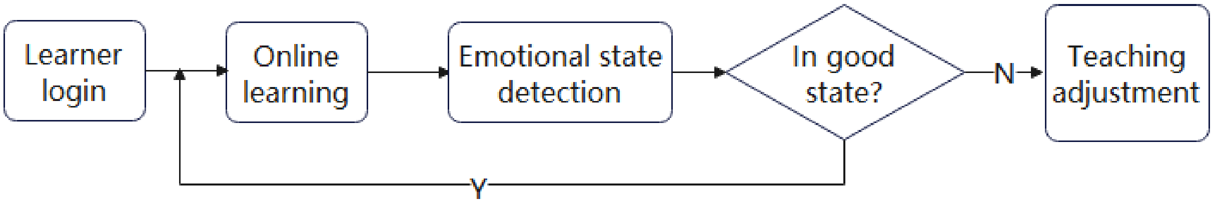


FIGURE 6. Design flow chart of online learning system.

of the time, it signifies a lack of active participation or potential disinterest, prompting the system to deliver timely reminders.

5. **Learning Preference:** A comparative analysis of students' emotional responses while learning different courses is performed. This analysis helps determine the types of courses that evoke happiness or surprise from students. This insight is subsequently utilized when suggesting future courses; the system can proactively recommend courses similar in nature or presented by the same instructor to align with students' positive learning experiences.

V. RESULT AND DISCUSSIONS

A. EXPERIMENTAL ENVIRONMENT

The facial expression recognition experiment has been conducted using the RAF-DB, FER2013, KDEF, and CK+ datasets to validate the efficacy of the proposed algorithm. RAF-DB stands as a substantial collection of real-world facial expressions, encompassing 29,672 images portraying diverse facial expressions sourced from the Internet. Within this dataset, 40 individuals annotated the assortment of 29,672 images, capturing a range of both basic and complex expressions. Meanwhile, the FER2013 dataset comprises 35,886 facial expression images, distributed among 28,708 Training images, 3,589 PublicTest images, and 3,589 PrivateTest images. Each image within this dataset is a grayscale representation with a fixed dimension of 48 × 48 pixels. Table 1 enumerates the seven distinct expressions included in the datasets.

To ascertain the durability of our algorithm across the two natural scene expression datasets, we utilize the training and validation sets from both datasets for model training. Through stochastic gradient descent, parameters are

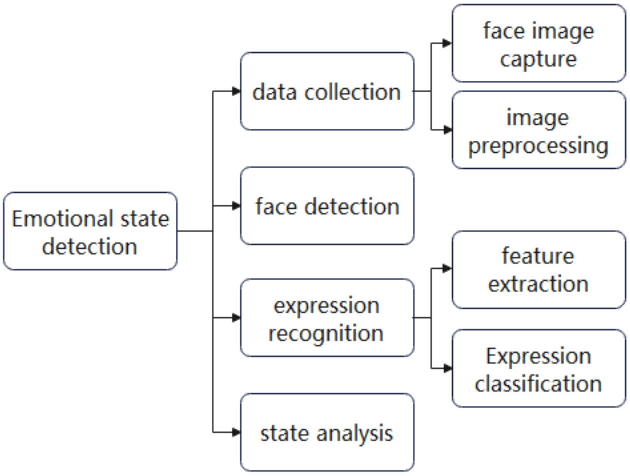


FIGURE 7. Emotional state detection module.

TABLE 1. Category information for RAF-DB and FER2013 datasets.

	RAF-DB		FER2013		
	Train	Test	Train	Public Test	PrivateTest
Anger	705	162	3995	467	491
Disgust	717	160	436	56	55
Fear	281	74	4097	496	528
Happy	4772	1185	7215	895	879
Sad	1982	478	4830	653	594
Surprise	1290	329	3171	415	416
Normal	2524	680	4956	607	626

iteratively adjusted to optimize recognition outcomes. Prior to commencement, parameters receive initialization, with the learning rate established at 0.0001 and a batch size of 16 selected to expedite gradient descent convergence and counteract issues like gradient vanishing or exploding.

The experimental arrangement involves a Windows 10 computer, employing the PyTorch deep learning framework to facilitate model training. All code-based experiments are carried out within the PyCharm development environment.

B. EXPERIMENTAL RESULTS ON FACIAL EXPRESSION RECOGNITION

1) FACIAL EXPRESSION RECOGNITION ON RAF-DB DATASET

The recognition confusion matrices for the seven facial expressions (happiness, disgust, anger, surprise, sadness, fear, and neutral) in both the initial and adjusted ResNet-50 models are presented in Figures 9 and 10, respectively. It's evident that there's a noticeable enhancement in the accuracy of ResNet-50 post-adjustment. This improvement can be attributed to the refined network's ability to better sustain the propagation of crucial information within the network structure, consequently enhancing facial expression recognition accuracy effectively.

Subsequent to the ResNet-50 adjustment, discernible enhancements are observed across all seven expressions in the RAF-DB dataset. The recognition accuracy for easily distinguishable expressions such as happiness and neutral has risen from 0.89 and 0.88 to 0.91 and 0.90, respectively. Similarly, recognition accuracy for relatively challenging expressions like fear and disgust has also shown improvement, escalating from 0.73 and 0.70 to 0.77 and 0.74, respectively.

In order to verify the reliability of the method proposed in this article, a comparison was made between the proposed method and the deep learning method that recently conducted facial expression recognition experiments on the RAF-DB dataset, and the accuracy was shown in Table 2.

Within the RAF-DB dataset, the distribution of expressions is uneven, notably for negative emotions like fear and disgust, which are comparatively less represented. Additionally, certain images exhibit compound expressions, introducing challenges during network training. When contrasted with alternative convolutional neural networks, this algorithm leverages ResNet-50 as its foundational architecture. This strategic choice proves effective in mitigating adverse outcomes linked to issues like vanishing gradients, often stemming from excessive network depth. As a result, it enables the extraction of more comprehensive feature information, culminating in an enhanced recognition performance.

2) FACIAL EXPRESSION RECOGNITION ON FER2013 DATASET

The recognition confusion matrices representing the seven facial expressions, both pre and post adjustment, for ResNet50 models on the public and private verification sets of FER2013 are depicted in Figures 11 and 12, respectively. Prior to fine-tuning, ResNet-50 exhibited recognition rates of 65.71% and 68.43% on the public and private verification sets, respectively. After fine-tuning, these rates experienced an enhancement, reaching 88.13 % and 73.43 % on the public and private verification sets, respectively. Evidently, the utilization of a fine-tuned ResNet-50 as the foundational network for expression recognition yields significant improvements in model performance.

When comparing expressions, it's notable that happiness and neutrality showcase distinct facial alterations, contributing to higher recognition rates. Specifically, on the private test, accuracy rates stand at 0.82 and 0.76, while on the public test, they amount to 0.93 and 0.95, respectively. Conversely, the variations among disgust, anger, fear, surprise, and sadness are less pronounced, resulting in comparatively higher misjudgment rates. These expressions yield accuracy rates of 0.65, 0.62, 0.70, 0.66, and 0.76 on the private test, and 0.81, 0.78, 0.86, 0.90, and 0.93 on the public test.

To ascertain the credibility of the proposed approach, a comparative analysis was conducted between the proposed method and a recently implemented deep learning technique in an expression recognition experiment using the FER2013 dataset. The resulting accuracy scores are detailed in Table 3.

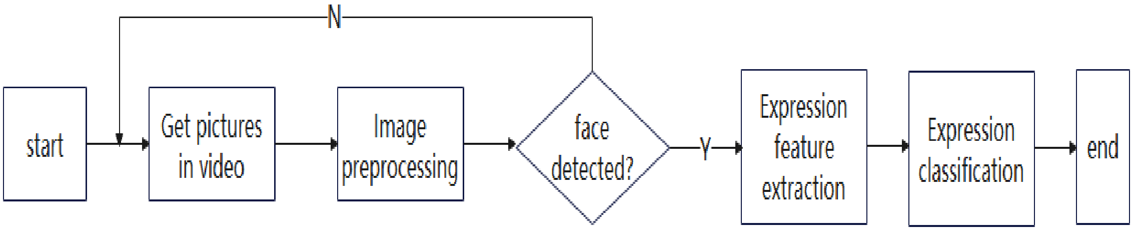


FIGURE 8. Expression recognition module.

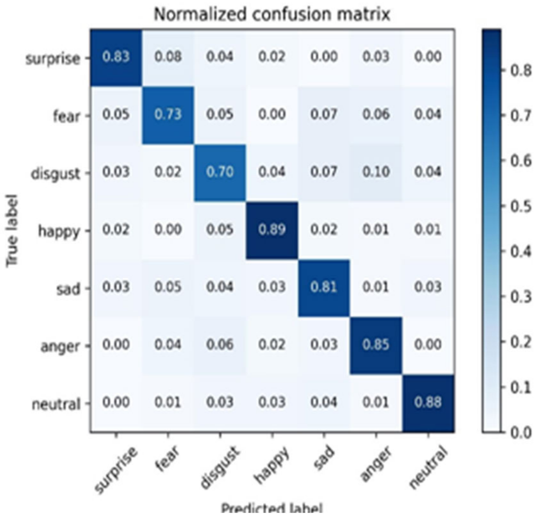


FIGURE 9. Confusion matrix of ResNet-50 identification results on RAF-DB.

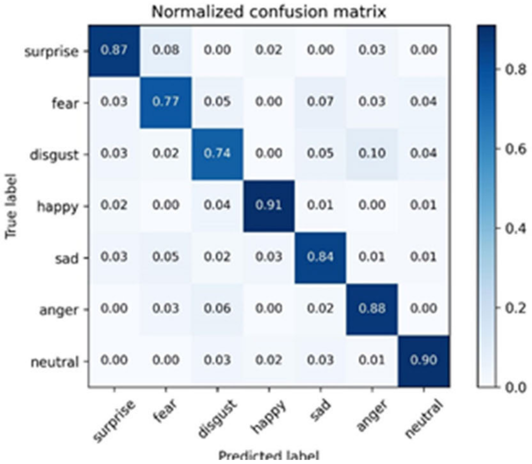


FIGURE 10. Confusion matrix of improved ResNet-50 identification results on RAF-DB.

TABLE 2. Accuracy of different models on RAF-DB dataset.

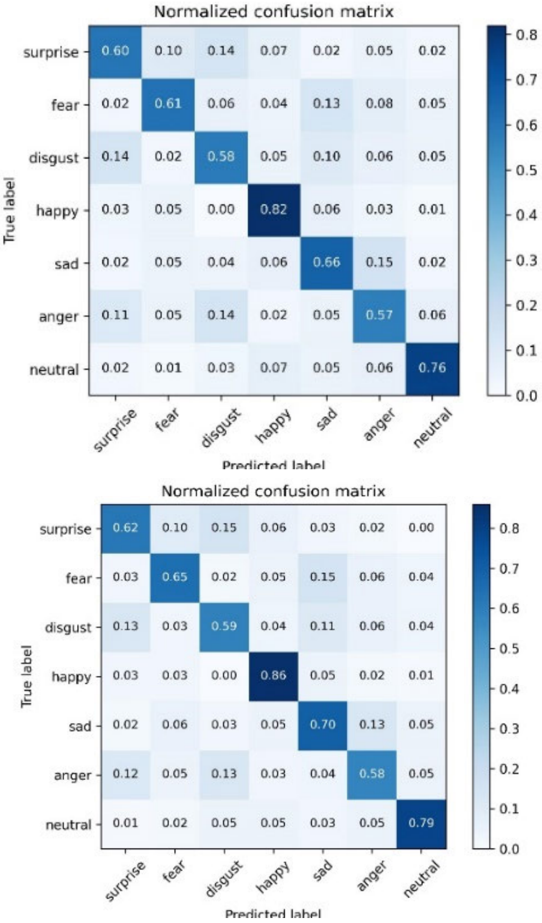
Dataset	Model	Accuracy
RAF-DB	ResNet	84.62 %
	[32]	84.75 %
	[13]	86.50 %
	Model in this paper	86.72 %

Given the minimal distinctions among akin samples within the FER2013 dataset, encompassing variations such as twisting, degree of twisting, and light intensity, the average recognition rate achieved by the proposed algorithm within the FER2013 dataset notably trails behind that within the RAF-DB dataset. The experimental findings indicate that the refined ResNet model surpasses alternative methods in enhancing expression classification accuracy. This, in turn, underscores the heightened feasibility and efficacy of the proposed approach.

3) CONDUCTING A COMPARATIVE ANALYSIS OF THE PROPOSED MODEL AGAINST STATE-OF-THE-ART TECHNIQUES

We conducted a series of experiments to assess the performance of our proposed model alongside other state-of-the-art methods, as detailed in Table 4. To gauge the model's resilience, we initially compared its performance to FER-2013. Our model outperformed the models of Arriaga et al. [35], J. Li et al. [36], Subramanian et al. [37], Borgalli et al. [38], Kunyoung Lee [20], Jiarui Zhong [21], and Hongmei Zhong [23], achieving accuracy improvements of 22.13%, 16.13%, 0.11%, 1.33%, 14.53%, 36.84%, and 11.33%, respectively. We also evaluated the resilience of our model using the CK+ dataset, where it demonstrated a noteworthy performance compared to state-of-the-art methods. In this evaluation, our proposed model attained an accuracy that was 13.58%, 6.88%, 15.99%, and 17.48% higher than that of Borgalli et al. [38], Bodapati et al. [39], Kunyoung Lee [20], and Hongmei Zhong [23], respectively. We proceeded to evaluate the proposed model using the KDEF dataset, where our model exhibited a superior accuracy, surpassing the method introduced by Sajjad et al. [40] by 2.87%. A more in-depth analysis of the model revealed that Haq et al. [41], Liu et al. [42], Kunyoung Lee [20], and Hongmei Zhong [23] achieved performances that were 2.57%, 9.37%, 11.35%, and 24.77% lower, respectively, compared to our proposed model.

The attainment of the highest accuracy by the proposed model primarily hinges on two key factors. Firstly, the incorporation of the residual module from ResNet-50 effectively addresses the issue of network degradation. Secondly, the model's depth, coupled with convolution, endows it with a



Normalized confusion matrix

True label \ Predicted label	surprise	fear	disgust	happy	sad	anger	neutral
surprise	0.60	0.10	0.14	0.07	0.02	0.05	0.02
fear	0.02	0.61	0.06	0.04	0.13	0.08	0.05
disgust	0.14	0.02	0.58	0.05	0.10	0.06	0.05
happy	0.03	0.05	0.00	0.82	0.06	0.03	0.01
sad	0.02	0.05	0.04	0.06	0.66	0.15	0.02
anger	0.11	0.05	0.14	0.02	0.05	0.57	0.06
neutral	0.02	0.01	0.03	0.07	0.05	0.06	0.76

Normalized confusion matrix

True label \ Predicted label	surprise	fear	disgust	happy	sad	anger	neutral
surprise	0.62	0.10	0.15	0.06	0.03	0.02	0.00
fear	0.03	0.65	0.02	0.05	0.15	0.06	0.04
disgust	0.13	0.03	0.59	0.04	0.11	0.06	0.04
happy	0.03	0.03	0.00	0.86	0.05	0.02	0.01
sad	0.02	0.06	0.03	0.05	0.70	0.13	0.05
anger	0.12	0.05	0.13	0.03	0.04	0.58	0.05
neutral	0.01	0.02	0.05	0.05	0.03	0.05	0.79

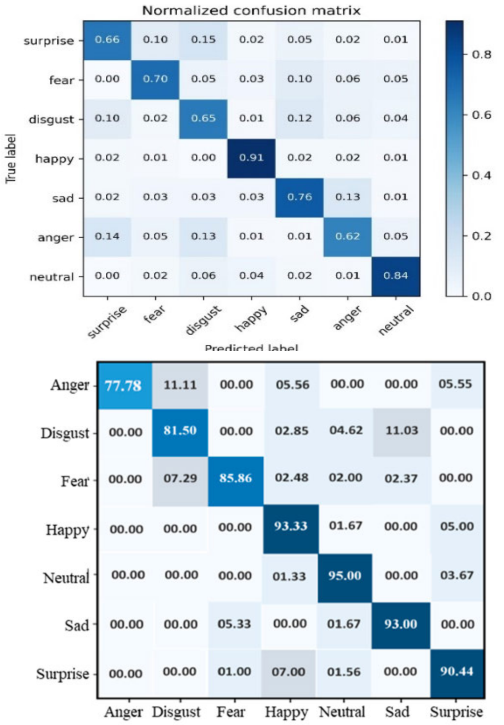
FIGURE 11. Confusion matrix of ResNet-50 identification results on private and public test of FER2013.

robust feature extraction capability and exceptional training prowess.

In this experiment, the ResNet-50 network has demonstrated favorable outcomes when compared to alternative methods. Our subsequent investigation will involve the exploration of various ResNet layers and several ResNet variants, including Wide Residual Network (WRN), ResNeXt, and MobileNet, for the study of facial emotion recognition. In our future endeavors, we will assess their performance.

4) THE TIME COMPLEXITY EVALUATION OF THE SUGGESTED MODEL ON GPU, CPU, AND RESOURCE-CONSTRAINED DEVICES

We conducted a real-time performance assessment of the proposed model to measure its processing time across various computing platforms, including GPU, CPU, and a resource-constrained device, the Jetson Nano. The Jetson Nano is a compact yet potent computer capable of running multiple CNNs simultaneously for diverse applications such as recognition, segmentation, object detection, and speech processing. It boasts 512 NVIDIA CUDA® cores in its GPU, a Quad-core ARM Cortex-A57 CPU, and 8 GB of memory.



Normalized confusion matrix

True label \ Predicted label	surprise	fear	disgust	happy	sad	anger	neutral
surprise	0.66	0.10	0.15	0.02	0.05	0.02	0.01
fear	0.00	0.70	0.05	0.03	0.10	0.06	0.05
disgust	0.10	0.02	0.65	0.01	0.12	0.06	0.04
happy	0.02	0.01	0.00	0.91	0.02	0.02	0.01
sad	0.02	0.03	0.03	0.03	0.76	0.13	0.01
anger	0.14	0.05	0.13	0.01	0.01	0.62	0.05
neutral	0.00	0.02	0.06	0.04	0.02	0.01	0.84

	Anger	Disgust	Fear	Happy	Neutral	Sad	Surprise
Anger	77.78	11.11	00.00	05.56	00.00	00.00	05.55
Disgust	00.00	81.50	00.00	02.85	04.62	11.03	00.00
Fear	00.00	07.29	85.86	02.48	02.00	02.37	00.00
Happy	00.00	00.00	00.00	93.33	01.67	00.00	05.00
Neutral	00.00	00.00	00.00	01.33	95.00	00.00	03.67
Sad	00.00	00.00	05.33	00.00	01.67	93.00	00.00
Surprise	00.00	00.00	01.00	07.00	01.56	00.00	90.44

FIGURE 12. Confusion matrix of improved ResNet-50 identification results on private and public test of FER2013.

TABLE 3. Accuracy of different models on FER2013 dataset.

Dataset	Model	Accuracy
FER2013	ResNet	57.92 %
	[33]	72.67 %
	[34]	73.00 %
	Model in this paper	88.13 %

The frames per second (fps) achieved by the proposed model on GPU, CPU, and Jetson Nano were 32, 14, and 19 s, respectively. These findings demonstrate that the proposed model exhibits significantly reduced time complexity, making it well-suited for real-world deployment scenarios.

In Table 5, we present a comparative analysis of our results on the CK+ and RAF-DB datasets. Our assessment was based on three key evaluation metrics: accuracy (%), run-time (in seconds), and CPU memory usage (average %). In the case of the CK+ dataset, our method achieved a test accuracy of 94.58%, which is only 2.45% lower than the highest accuracy achieved by the PG-CNN method [43]. Notably, our method demonstrated an impressive run-time of just 61 seconds.

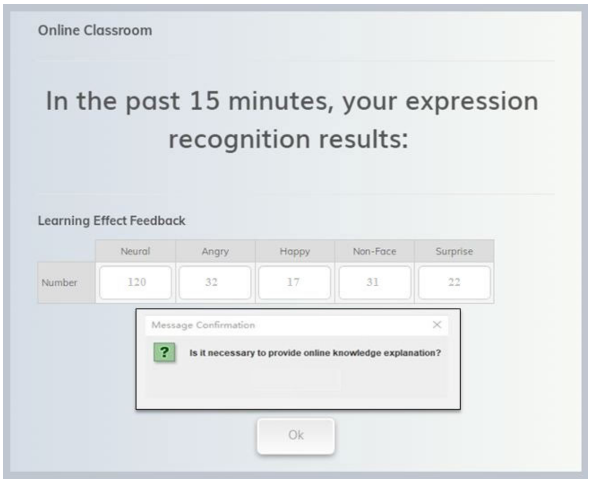
Compared to recent approaches, our proposed method strikes a balance between accuracy and run-time, a crucial consideration for the feasibility of deploying deep learning techniques on edge devices for real-world applications. To summarize, on the CK+ dataset, our proposed method not only delivered competitive accuracy but also showcased computational efficiency. In the context of the RAF-DB

TABLE 4. Assessing the performance of the proposed model in comparison to the state-of-the-art method across three benchmark datasets.

Dataset	References	Methods	Average Accuracy
FER-2013	Arriaga et al. [35]	Mini-Xception	66 %
	J. Li et al. [36]	CNN with Transfer Learning	72 %
	Subramanian et al. [37]	Three Layer CNN architecture	88.02 %
	Borgalli et al. [38]	Six Layer CNN architecture	86.8 %
	Kunyoung Lee [20]	MobileNetV2	73.6 %
	Jiarui Zhong [21]	NGO-BILSTM	51.29 %
	Hongmei Zhong [23]	CBSA + CSPA	76.8 %
	The proposed model	proposed	**88.13** %
CK+	Borgalli et al. [38]	Six Layer CNN architecture	81 %
	Bodapati et al. [39]	InceptionResNetV2	87.7 %
	Kunyoung Lee [20]	MobileNetV2	78.59%
	Hongmei Zhong [23]	CBSA + CSPA	77.1 %
	The proposed model	proposed	**94.58** %
KDEF	Sajjad et al. [40]	Fine-tuned AlexNet	93.4 %
	Haq et al. [41]	CNN with Transfer Learning	93.7 %
	Liu et al. [42]	Multi-channel features	86.9 %
	Kunyoung Lee [20]	MobileNetV2	84.92 %
	Hongmei Zhong [23]	CBSA + CSPA	71.5 %
	The proposed model	proposed	**96.27** %

TABLE 5. Presents a comparison with the latest state-of-the-art results obtained from the CK+ and RAF-DB datasets. The run-time refers to the duration required for recognizing 1492 images in the CK+ dataset and 2920 images in the RAF-DB dataset, respectively.

Dataset	Methods	Run Time (s)	Accuracy (%)	CPU memory (%)
CK+	PG-CNN [43]	491	**97.03**	82.6
	gACNN [44]	492	96.40	67.3
	eXnet [45]	273	95.81	61.7
	The proposed model	**61**	94.58	**41.9**
RAF-DB	ResiDen [46]	656	76.54	63.1
	SCN [47]	522	**87.03**	78.6
	gACNN [44]	568	85.07	64.5
	Light-CNN [48]	163	77.23	51.5
	The proposed model	**115**	86.72	**38.4**



Online Classroom

In the past 15 minutes, your expression recognition results:

Learning Effect Feedback

	Neural	Angry	Happy	Non-Face	Surprise
Number	120	32	17	31	22

Message Confirmation

? Is it necessary to provide online knowledge explanation?

FIGURE 13. Learning effect feedback.

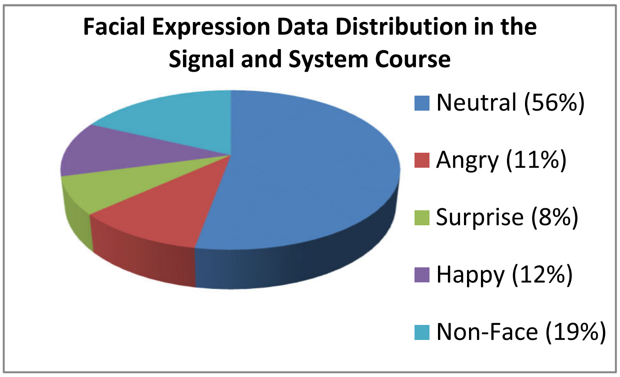


FIGURE 14. Distribution of facial expression data in Signal and system course.

dataset, our proposed method achieved a test accuracy of 86.72%, which is just 0.31% below the state-of-the-art result obtained by SCN [47]. However, what sets our method apart is its significantly faster run-time, clocking in at only about one-fourth of the time required by SCN [47]. Although our method exhibits somewhat reduced robustness when applied to the in-the-wild RAF-DB dataset compared to the controlled in-the-lab CK+ dataset, it's important to highlight that our method maintains a crucial advantage. For the 2920 test images, our method completes the recognition process in just 115 seconds, whereas other methods achieving accuracy above 85% typically demand approximately 400 seconds or more. This not only underscores our method's suitability for real-time applications but also its compatibility with edge devices.

C. EFFECT OF ONLINE LEARNING SYSTEM

The learning process yields expression recognition outcomes over a defined time span in the past, subsequently furnishing feedback and recommendations. This sequence is depicted in Fig. 13. Fig.14. Shows a pie chart showing students' expression recognition results.

VI. CONCLUSION

The essential findings of the experimental work should be outlined. Emphasis should be placed on the contribution of this study to both the scientific community and its potential economic implications.

Addressing the challenges posed by the current online learning systems, which suffer from "emotional loss" and a lack of feedback on learning outcomes, this paper leverages facial attributes to analyze and derive learning feedback. It establishes an online learning system grounded in facial expression recognition. The proposed methodology leverages the improved residual module within ResNet-50 to classify students' facial expressions, thereby progressively enhancing teaching and learning dynamics. Comparative assessments against multiple existing methods substantiate that our approach attains a notably high accuracy rate.

In this paper, we have improved the accuracy of facial expression recognition by employing ResNet-50 for efficient feature extraction. Our proposed model has demonstrated notably superior results when compared to alternative approaches. Through the evaluation of both quantitative and qualitative aspects using three distinct datasets, we have established the effectiveness of our proposed method in enhancing E-learning. The achievement of the highest accuracy by our model primarily relies on two crucial factors. Firstly, the integration of the residual module from ResNet-50 effectively addresses the issue of network degradation. Secondly, the model's depth, combined with convolution, equips it with a robust capability for feature extraction and exceptional training proficiency.

The real-time application of expression recognition technology to monitor students' facial expressions furnishes educators with visual classroom insights. This empowers teachers to gauge students' comprehension levels, promptly adjust teaching content and online resources, stimulate students' engagement, and elevate the overall quality of online instruction.

Nonetheless, while this paper conducts foundational research on monitoring online learning statuses through expression analysis, the method and system design presented herein exhibit certain limitations. Consequently, the following challenges and avenues for future research emerge:

1) **Image Recognition Challenges:** In the process of gathering images from the online learning environment, learners' faces and expressions may not be accurately identified due to a variety of factors. These factors should be meticulously examined to enhance the congruence of recognition results with actual circumstances.
2) **Complex Student States:** The current classification of students' states might be oversimplified. It's imperative to delve into more intricate expression data to construct and refine a broader spectrum of potential student states within the classroom.

CONFLICTS OF INTEREST

The authors declare no conflict of interest. The authors have no competing interests to declare that are relevant to the content of this article. The submitted work is original and has not been submitted to another journal for simultaneous consideration. The manuscript is not published elsewhere in any form or language.

Author Contributions:

Conceptualization: M.A., A.G., and I.S.; Methodology: M.A., A.G., and I.S.; Formal analysis and investigation: M.A., A.G., and I.S.; Writing—original draft preparation: M.A., A.G., and I.S.; Writing—review and editing: M.A., and A.G.; Resources: A.G., and I.S.; Supervision: A.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Declaration

The authors declare that they have no conflicts of interest to report regarding the present study.

REFERENCES

[1] A. V. Savchenko, L. V. Savchenko, and I. Makarov, "Classifying emotions and engagement in online learning based on a single facial expression recognition neural network," *IEEE Trans. Affect. Comput.*, vol. 13, no. 4, pp. 2132–2143, Oct. 2022, doi: 10.1109/TAFFC.2022.3188390.

[2] P. K. Sidhu, A. Kapoor, Y. Solanki, P. Singh, and D. Sehgal, "Deep learning based emotion detection in an online class," in *Proc. IEEE Delhi Sect. Conf. (DELCON)*, New Delhi, India, Feb. 2022, pp. 1–6, doi: 10.1109/DELCON54057.2022.9752940.

[3] H. Hu, E. Real, K. Takamiya, M.-G. Kang, J. Ledoux, R. L. Huganir, and R. Malinow, "Emotion enhances learning via norepinephrine regulation of AMPA-receptor trafficking," *Cell*, vol. 131, no. 1, pp. 160–173, Oct. 2007, doi: 10.1016/j.cell.2007.09.017.

[4] S. D'Mello and A. Graesser, "Dynamics of affective states during complex learning," *Learn. Instruct.*, vol. 22, no. 2, pp. 145–157, Apr. 2012, doi: 10.1016/j.learninstruc.2011.10.001.

[5] M. Jagadeesh and B. Baranidharan, "Facial expression recognition of online learners from real-time videos using a novel deep learning model," *Multimedia Syst.*, vol. 28, no. 6, pp. 2285–2305, Dec. 2022, doi: 10.1007/S00530-022-00957-Z.

[6] F. Yuan and Y. Nie, "Online classroom teaching quality evaluation system based on facial feature recognition," *Scientific Program.*, vol. 2021, pp. 1–10, Dec. 2021, doi: 10.1155/2021/7374846.

[7] Z. Zhan, "An intelligent agent-based emotional and cognitive recognition model for distance learners: Coupling supported by eye movement tracking and expression recognition technology," *Modern Distance Educ. Res.*, vol. 5, pp. 100–105, Jan. 2013, doi: 10.3969/j.issn.1009-5195.2013.05.013.

[8] N. Ouherrou, O. Elhammoumi, F. Benmarrakchi, and J. El Kafi, "Comparative study on emotions analysis from facial expressions in children with and without learning disabilities in virtual learning environment," *Educ. Inf. Technol.*, vol. 24, no. 2, pp. 1777–1792, Mar. 2019, doi: 10.1007/s10639-018-09852-5.

[9] X. Yang and Z. Shang, "Facial expression recognition based on improved AlexNet," *Laser Optoelectronics Prog.*, vol. 57, no. 14, pp. 235–242, 2020, doi: 10.3788/LOP57.141026.

[10] Z. Cui, J. Pi, Y. Chen, J. Yang, Y. Xian, Z. Wu, L. Zhao, S. Zeng, and J. Lv, "Facial expression recognition combined with improved VGGNet and focal loss," *Comput. Eng. Appl.*, vol. 57, no. 19, pp. 171–178, 2021, doi: 10.3778/j.issn.1002-8331.2007-0492.

[11] A. Caroppo, A. Leone, and P. Siciliano, "Comparison between deep learning models and traditional machine learning approaches for facial expression recognition in ageing adults," *J. Comput. Sci. Technol.*, vol. 35, no. 5, pp. 1127–1146, Oct. 2020, doi: 10.1007/s11390-020-9665-4.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2016, pp. 770–778.

[13] Z. Wang, F. Zeng, S. Liu, and B. Zeng, "OAENet: Oriented attention ensemble for accurate facial expression recognition," *Pattern Recognit.*, vol. 112, Apr. 2021, Art. no. 107694, doi: 10.1016/j.patcog.2020.107694.

[14] M. Neji and M. B. Ammar, "Agent-based collaborative affective e-learning framework," *Electron. J. e-Learn.*, vol. 5, no. 2, pp. 123–134, 2007.

[15] K. Bahreini, R. Nadolski, and W. Westera, "FILTWAM—A framework for online affective computing in serious games," *Proc. Comput. Sci.*, vol. 15, pp. 45–52, Jan. 2012, doi: 10.1016/j.procs.2012.10.057.

[16] N. Zhou, R. Liang, and W. Shi, "A lightweight convolutional neural network for real-time facial expression detection," *IEEE Access*, vol. 9, pp. 5573–5584, 2021, doi: 10.1109/ACCESS.2020.3046715.

[17] W. C. de Melo, E. Granger, and A. Hadid, "A deep multiscale spatiotemporal network for assessing depression from facial dynamics," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1581–1592, Jul. 2022, doi: 10.1109/TAFFC.2020.3021755.

[18] L. Liang, C. Lang, Y. Li, S. Feng, and J. Zhao, "Fine-grained facial expression recognition in the wild," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 482–494, 2021, doi: 10.1109/TIFS.2020.3007327.

[19] Y. ELsayed, A. ELSayed, and M. A. Abdou, "An automatic improved facial expression recognition for masked faces," *Neural Comput. Appl.*, vol. 35, no. 20, pp. 14963–14972, Jul. 2023, doi: 10.1007/s00521-023-08498-w.

[20] K. Lee, S. Kim, and E. C. Lee, "Fast and accurate facial expression image classification and regression method based on knowledge distillation," *Appl. Sci.*, vol. 13, no. 11, pp. 1–14, 2023, doi: 10.3390/app13116409.

[21] J. Zhong, T. Chen, and L. Yi, "Face expression recognition based on NGO-BILSTM model," *Frontiers Neurorobotics*, vol. 17, pp. 1–10, Mar. 2023, doi: 10.3389/fnbot.2023.1155038.

[22] A. L. Cîrneanu, D. Popescu, and D. Iordache, "New trends in emotion recognition using image analysis by neural networks," *Sensors*, vol. 23, no. 16, p. 7092, 2023, doi: 10.3390/s23167092.

[23] H. Zhong, T. Han, W. Xia, Y. Tian, and L. Wu, "Research on real-time teachers' facial expression recognition based on YOLOv5 and attention mechanisms," *EURASIP J. Adv. Signal Process.*, vol. 2023, no. 1, pp. 1–15, May 2023, doi: 10.1186/s13634-023-01019-w.

[24] G. A. R. Kumar, R. K. Kumar, and G. Sanyal, "Facial emotion analysis using deep convolution neural network," in *Proc. Int. Conf. Signal Process. Commun. (ICSPC)*, Coimbatore, India, Jul. 2017, pp. 369–374, doi: 10.1109/CSPC.2017.8305872.

[25] 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 *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, San Francisco, CA, USA, Jul. 2010, pp. 94–101, doi: 10.1109/CVPRW.2010.5543262.

[26] S. Barrett, F. Weimer, and J. Cosmas, "Virtual eye region: Development of a realistic model to convey emotion," *Heliyon*, vol. 5, no. 12, Dec. 2019, Art. no. e02778.

[27] 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, pp. 1–15, 2023, doi: 10.3390/electronics12173595.

[28] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–15.

[29] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 286–3013.

[30] M. Aly and N. S. Alotaibi, "A novel deep learning model to detect COVID-19 based on wavelet features extracted from mel-scale spectrogram of patients' cough and breathing sounds," *Informat. Med. Unlocked*, vol. 32, pp. 1–11, 2022.

[31] S. Yu, S. Jin, J. Peng, H. Liu, and Y. He, "Application of a new deep learning method with CBAM in clothing image classification," in *Proc. IEEE Int. Conf. Emergency Sci. Inf. Technol. (ICESIT)*, Chongqing, China, Jul. 2021, pp. 364–368, doi: 10.1109/ICESIT53460.2021.9696783.

[32] H. Zhang, W. Su, J. Yu, and Z. Wang, "Identity–expression dual branch network for facial expression recognition," *IEEE Trans. Cognit. Develop. Syst.*, vol. 13, no. 4, pp. 898–911, Dec. 2021, doi: 10.1109/TCDS.2020.3034807.

[33] W. Xie, L. Shen, and J. Duan, "Adaptive weighting of handcrafted feature losses for facial expression recognition," *IEEE Trans. Cybern.*, vol. 51, no. 5, pp. 2787–2800, May 2021, doi: 10.1109/TCYB.2019.2925095.

[34] J. Chen and Y. Xu, "Expression recognition based on convolution residual network of attention pyramid," *Comput. Eng. Appl.*, vol. 58, no. 22, pp. 123–131, 2022, doi: 10.3778/j.issn.1002-8331.2104-0111.

[35] O. Arriaga, M. Valdenegro-Toro, and P. Plöger, "Real-time convolutional neural networks for emotion and gender classification," 2017, *arXiv:1710.07557*.

[36] J. Li, Y. Wang, J. See, and W. Liu, "Micro-expression recognition based on 3D flow convolutional neural network," *Pattern Anal. Appl.*, vol. 22, no. 4, pp. 1331–1339, Nov. 2019, doi: 10.1007/s10044-018-0757-5.

[37] R. R. Subramanian, C. S. Niharika, D. U. Rani, P. Pavani, and K. P. L. Syamala, "Design and evaluation of a deep learning algorithm for emotion recognition," in *Proc. 5th Int. Conf. Intell. Comput. Control Syst. (ICICCS)*, Madurai, India, May 2021, pp. 984–988, doi: 10.1109/ICICCS51141.2021.9432336.

[38] M. R. Appasaheb Borgalli and D. S. Surve, "Deep learning for facial emotion recognition using custom CNN architecture," *J. Phys., Conf. Ser.*, vol. 2236, no. 1, Mar. 2022, Art. no. 012004, doi: 10.1088/1742-6596/2236/1/012004.

[39] J. D. Bodapati, D. S. B. Naik, B. Suvarna, and V. Naralasetti, "A deep learning framework with cross pooled soft attention for facial expression recognition," *J. Inst. Eng. India, Ser. B*, vol. 103, no. 5, pp. 1395–1405, Oct. 2022, doi: 10.1007/s40031-022-00746-2.

[40] M. Sajjad, S. Zahir, A. Ullah, Z. Akhtar, and K. Muhammad, "Human behavior understanding in big multimedia data using CNN based facial expression recognition," *Mobile Netw. Appl.*, vol. 25, no. 4, pp. 1611–1621, Aug. 2020, doi: 10.1007/s11036-019-01366-9.

[41] I. Ul Haq, A. Ullah, K. Muhammad, M. Y. Lee, and S. W. Baik, "Personalized movie summarization using deep CNN-assisted facial expression recognition," *Complexity*, vol. 2019, pp. 1–10, May 2019, doi: 10.1155/2019/3581419.

[42] Y. Liu, J. Zeng, S. Shan, and Z. Zheng, "Multi-channel pose-aware convolution neural networks for multi-view facial expression recognition," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Xi'an, China, 2018, pp. 458–465, doi: 10.1109/FG.2018.00074.

[43] Y. Li, J. Zeng, S. Shan, and X. Chen, "Patch-gated CNN for occlusion-aware facial expression recognition," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Beijing, China, 2018, pp. 2209–2214, doi: 10.1109/ICPR.2018.8545853.

[44] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion aware facial expression recognition using CNN with attention mechanism," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2439–2450, May 2019, doi: 10.1109/TIP.2018.2886767.

[45] 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, pp. 1–12, 2020, doi: 10.3390/s20041087.

[46] S. Jyoti, G. Sharma, and A. Dhall, ‘‘Expression empowered ResiDen network for facial action unit detection,’’ in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Lille, France, May 2019, pp. 1–8, doi: 10.1109/FG.2019.8756580.

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

[48] J. Shao and Y. Qian, ‘‘Three convolutional neural network models for facial expression recognition in the wild,’’ *Neurocomputing*, vol. 355, pp. 82–92, Aug. 2019, doi: 10.1016/j.neucom.2019.05.005.

MOHAMMED ALY received the B.S. degree from the University of Zagazig, Zagazig, Egypt, in 2012, the M.S. degree in computer science from the Faculty of Science, Zagazig University, in 2017, and the Ph.D. degree in computer science from the Faculty of Science, Al-Azhar University, Egypt, in 2020. He is currently an Associate Professor with the College of Artificial Intelligent, Egyptian Russian University (ERU). His research interests include artificial intelligence, image processing, video processing, classification, detection, machine learning, deep learning, and the Internet of Things (IoT).

ABDULLATIF GHALLAB is currently the Director of Graduate Studies and Scientific Research with the Faculty of Computing and Information Technology. During the past ten years, about 28 master’s and Ph.D. students completed their theses or projects under the supervision of Dr. Ghallab. He also participated in the evaluation committees for more than 63 theses and research proposals. His research interests include computational intelligence, machine learning, social analytics, business intelligence, IT, enterprise systems, knowledge management, e-business, and mobile HCI.

ISLAM S. FATHI received the B.Sc. and M.Sc. degrees in mathematics and computer sciences from the Faculty of Science, Zagazig University, Egypt, in 2013 and 2019, respectively, and the Ph.D. degree in computer science from the Faculty of Science, Suez Canal University, Egypt, in 2023. He is currently a Lecturer in computer science with the Department of Information Systems, Al Alson Academy, Cairo, Egypt. His research interests include signal processing, metaheuristic optimization, bioinformatics, and the Internet of Things.