Ain Shams Engineering Journal 16 (2025) 103584

Contents lists available at ScienceDirect

Ain Shams Engineering Journal

journal homepage: https://www.sciencedirect.com

Full Length Article

A multi-scale simplicial transformer with graph attention for facial emotion recognition

Samia Nawaz Yousafzai [a,1], Inzamam Mashood Nasir [b], Oumaima Saidani [c], Refka Ghodhbani [d,*], Yeonghyeon Gu [e,*], Muhammad Syafrudin [e,*], Norma Latif Fitriyani [e,1]

[a] Department of Computer Science, HITEC University Taxila, Taxila, 47080, Pakistan
[b] Faculty of Informatics, Kaunas University of Technology, Kaunas, 51368, Lithuania
[c] Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia
[d] Center for Scientific Research and Entrepreneurship, Northern Border University, Arar, 73213, Saudi Arabia
[e] Department of Artificial Intelligence and Data Science, Sejong University, Seoul, 05006, Republic of Korea

ARTICLE INFO

Dataset link: https://github.com/Samia-Nawaz/Facial-Emotion-Recognition

Keywords:
Facial expression recognition
Face detection
Graph attention network
Simplicial transformer
Hybrid adaptive attention
Explainable AI

ABSTRACT

Facial Emotion Recognition (FER) plays a vital role in human-computer interaction and affective computing, facing challenges like obstructed views and varying facial poses. Our approach employs a graph-based FER framework integrating multi-scale feature extraction with adaptive attention mechanisms for accurate emotion detection. Initially, YOLOv8 detects faces, enabling the creation of multi-scale graphs to analyze spatial relationships among features. A hybrid adaptive attention mechanism sharpens these features before processing them by a simplicial transformer network for dependency capture. Using a graph attention network enhances edge weighting, thereby improving recognition performance. The proposed model is evaluated on two benchmark datasets namely AffectNet and FER2013 achieving accuracy of 81.84% and 90.40%, respectively. On occlusion and pose AffectNet dataset, the model demonstrates notable accuracy improvements of 3.7% and 4.2%, respectively, over the strongest baseline. Futhermore, cross-dataset validation is conducted with highest performance of 98.54% accuracy by combining (AffectNet and FER2013) for training and testing on additional CK+ dataset. Across these datasets, statistical significance is confirmed through paired t-tests and Wilcoxon signed-rank tests, with p-values consistently below 0.05, validating the robustness of performance gains. Visualizations using Grad-CAM and t-SNE further validate the model's discriminative power and focus on expressive regions. These results demonstrate strong generalization and practical applicability of the proposed approach in real-world FER scenarios.

1. Introduction

Facial expression, a crucial biological aspect of emotional cognition, is one of the most basic and direct methods for conveying human emotions. It is also a potent tool for understanding and communicating emotional states. Researchers are increasingly intrigued by the impact of expression-based emotional intelligence on the advancement of artificial intelligence. Automatic facial expression recognition (FER) has numerous applications in various domains, such as human-computer interaction, psychological evaluation, medical surveillance, and public safety [1–3]. As a result, investigating facial expression recognition has attracted greater attention from researchers, leading to an ongoing development of related studies.

In controlled laboratory contexts, advancements in FER have led to exceptional recognition performance on small-scale, single-background, non-occlusion, and non-pose variant expression datasets, such as MMI [4], Extended Cohn-Kanade (CK+) [5], and Oulu-CASIA [6]. The recognition of expressions in these contexts is significantly complicated by

* Corresponding authors.
E-mail addresses: samia.nawaz@hitecuni.edu.pk (S.N. Yousafzai), inzamam.nasir@ktu.edu (I.M. Nasir), ocsaidani@pnu.edu.sa (I. Saidani), refka.ghodhbani@nbu.edu.sa (R. Ghodhbani), yhgu@sejong.ac.kr (Y. Gu), udin@sejong.ac.kr (M. Syafrudin), norma@sejong.ac.kr (N.L. Fitriyani).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.asej.2025.103584
Received 9 March 2025; Received in revised form 31 May 2025; Accepted 16 June 2025
Available online 2 July 2025

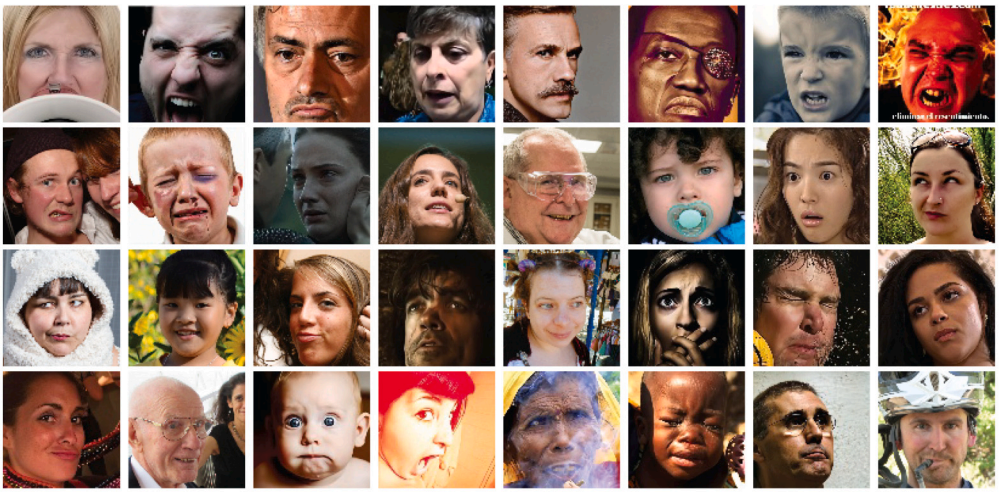

Fig. 1. Sample images taken from selected FER datasets.

the numerous occlusions, positions, nuanced expression fluctuations, lighting conditions, and image quality endemic in real-world settings, as illustrated in Fig. 1. The efficacy of recognition on extensive facial expression datasets in natural contexts, such as Real-world Affective Faces Database (RAF-DB) [7], FER2013 [8], and AffectNet [9], is still significantly improvable. As a result, expression recognition in real-world environments has become a primary focus of contemporary research.

In the initial stages of FER research, conventional machine learning techniques are primarily employed to extract engineering features from tiny, controlled expression datasets [10,11], which leads to a limitation in recognition efficacy. End-to-end network models considerably enhance the efficiency of expression identification. At the same time, deep learning-based feature extraction can more comprehensively capture intricate expression data as the demand for large-scale authentic expression applications increases. The performance of the wild FER task has been consistently improved by an increasing number of deep learning models [12,13]. The emphasis on core regions has been further investigated in wild FER [14,15] as attention mechanisms are increasingly employed in various computational tasks [16–18]. To optimize expressions' inter- and intraclass distribution, particular academics have devised island loss [19] and center loss [20], thereby improving the model's learning effectiveness. However, the following areas require further attention: Substantial challenges are presented by the prevalence of a multitude of occlusions, poses, and subtle expression variations across classes of actual facial expressions. Most contemporary methods do not provide a detailed comprehension of the overall expression and adequate feature extraction capabilities for regions of interest, which rely on a specific feature extraction and representation space. Consequently, the robustness of recognition is diminished. Additionally, the final recognition performance may be further compromised if the labeling bias results from subjective and objective factors and the significant interclass similarity and intraclass variability in wild expression datasets are not optimally addressed during model training. The main contributions are given as follows:

This study addresses critical challenges in facial emotion recognition related to occlusion, pose variation, and subtle expression shifts. A Multi-Scale Adaptive Graph (MSAG) module is proposed to capture fine-grained spatial dependencies between facial regions, improving upon the limited relational modeling of CNN-based methods. A Hybrid Adaptive Attention (HAA) mechanism is introduced to emphasize emotion-relevant regions by combining spatial and channel-level attention, addressing the uniform treatment of features in prior approaches. Additionally, a Multi-Scale Simplicial Transformer (MSST) integrates hierarchical transformer representations with topological attention, bridging the gap between semantic abstraction and structural awareness. These components collectively enhance robustness and cross-dataset generalization in complex real-world conditions.

The article is structured as follows: Section 2 reviews existing methods and their limitations. The detailed description of each module of the proposed methodology is given in Section 3. Experimental sections, qualitative assessments, and method comparison studies are provided in Section 4. Section 5 delivers essential findings and details limitations and possible approaches for further study.

2. Literature review

In the past few years, there has been a significant advancement in FER based on deep learning. Relative Uncertainty Learning (RUL) represents a new method for FER that uses relative uncertainty modeling through feature analysis combined with mixup techniques instead of relying on conventional Gaussian distribution assumptions. RUL performs better than existing uncertainty-learning models by delivering 60.66% accuracy for AffectNet alongside 73.75% accuracy for FER2013 [21]. The self-learning framework Point Adversarial Self Mining (PASM) employs point adversarial mining to enhance feature retrieval for FER under situations of blocked vision and role-evolving poses. The PASM obtained 88.68% accuracy on RAF-DB while reaching 73.59% accuracy on FER2013 [22]. The identity and Pose Disentangled Facial Expression Recognition (IPD-FER) model utilizes pre-trained face recognition models to encode expression- and pose-invariant identification information. Labeled data optimizes posture and expression encoders. Adversarial learning and feature disentanglement build neutral and expressive faces and compare them to identity and stance. It scored 99.28% on CK+, 88.89% on RAF-DB, 88.42% on FERPlus, 62.23% on AffectNet, and 87.66% on high pose-variant situations [23]. Adaptive Multi-layer Perceptual Attention Network (AMP-Net) improves FER performance through global and local perception and attention-based modules, which produce refined emotional features that overcome optical obstructions and modifications in facial position. AMP-Net demonstrated results of 89.25% on RAF-DB along with 64.54% on AffectNet-7 and achieved 74.48% on FER-2013 [24].

A Light Attention Embedding Network (LAENet-SA) strengthens FER systems by adding spatial attention components to various network levels, enhancing feature recognition in tough real-world challenges. LAENet-SA reached a performance rate of 97.86% on CK+ and 61.22% on AffectNet-8 while achieving 64.09% on AffectNet-7 and 68.25% on FED-RO surpassing other attention-based FER models [25]. A three-stage combined feature selection approach using CFS, RFE, and XGBoost boosts facial expression recognition. The model delivers DISFA results at 94.68% while it reaches 98.82% accuracy on CK+ along with 95.8% on DISFA+ and 97.7% on JAFFE and 66.1% on FER2013 yet CK+ and

JAFFE led to better performance compared to FER2013 [26]. Facial Geometry Enhanced Network (FGENet) represents a lightweight facial expression recognition model that employs three enhancement modules comprising RFE, GSConv, and DEA to optimize features and attention processing. The model operates successfully on FER2013 with 70.49% accuracy alongside 97.89% accuracy on CK+ and 86.72% accuracy on RAF database [27].

The research suggests using Visual Transformers and Feature Fusion (VTFF) to improve uncontrolled environment FER performance. Attention Selective Fusion (ASF) merges CNN local and global features before Transformer-based self-attention modeling. The model scored 88.14% and 88.81% on RAF-DB and FERPlus and 61.85% on AffectNet [28]. Cross-fusion dual-attention network (CF-DAN) solve uses dual-attention processing, a new C2 activation function, and self-attention knowledge distillation to refine local characteristics, globally acquire information, minimize data superfluousness, and improve model generalization. The model surpasses prior techniques with 92.78%, 92.02%, and 63.58% recognition accuracy on RAF-DB, FERPlus, and AffectNet datasets [29]. Vision Transformers are used to build a facial expression recognition system using HLA-ViT and hybrid local attention. The model creates two streams to extract global and hybrid local contextual data before merging them in decision-making. The proposed technique performs 90.45% on RAF-DB, 90.13% on FERPlus, and 65.07% on AffectNet [30]. Efficient-SwishNet, a facial emotion recognition model, employs EfficientNet-b0 architecture updated with Swish activation to improve feature extraction and classification speed. Efficient-SwishNet classifies 100% on CK+, 91.23% on JAFFE, 84.79% on KDEF, and 72.91% on FER-2013 [31]. The authors presented FER-net, a CNN-based model for efficient facial expression detection based on softmax classifier. Over seven emotional classes, the model was tested on five benchmark datasets including FER2013 and CK+, obtaining accuracies 78.9% and 97.8% respectively [32]. FLEPNet, a texture-based ensemble DCNN with multi-scale convolutional and residual blocks, addresses FER concerns such overfitting and intra-class variability. Raw image data, homomorphic filtering for illumination normalization, and four texture descriptors strengthen features. Upon evaluating five benchmark datasets, FLEP-Net achieved an impressive accuracy of 98.94% on CK+ and 80.72% on FER2013 [33].

Research FER faces multiple obstacles because deep learning models cause high computational complexity while being sensitive to occlusions, head pose variances, and their limited capabilities in unstable environments [30]. Most existing models depend on CNNs, but these networks face challenges with limited receptive field capacity and need extensive computational capabilities [27]. Nevertheless, research-adopted feature selection methods encounter difficulties because of information loss and performance reduction, mainly in high-dimensional settings [26]. Stability problems occur when using ReLU activation because the function causes neuron deactivation, which results in impaired learning capability [29].

In contrast to prior CNN or transformer-based FER models, the proposed framework introduces a tightly integrated design that unifies spatial graph modeling, attention-based refinement, and topological abstraction. Most existing methods either rely on local convolutions without capturing inter-region relationships or apply transformer layers without preserving the structural geometry of facial landmarks. The present work addresses these gaps through three core innovations: a Multi-Scale Adaptive Graph (MSAG) that captures directional spatial dependencies via GATv2, a Hybrid Adaptive Attention (HAA) mechanism that assigns dynamic importance to spatial and channel dimensions, and a Multi-Scale Simplicial Transformer (MSST) that preserves higher-order topological structures while aggregating global context. This integrated architecture enables robust emotion recognition even under pose variation, occlusion, and expression ambiguity—limitations often unaddressed by previous models.

3. Proposed methodology

The proposed facial emotion recognition framework is composed of four core components arranged in a systematic pipeline. First, YOLOv8 is employed for detecting facial regions of interest from raw input images. These regions are then transformed into multi-scale graphs using the Multi-Scale Adaptive Graph (MSAG) module, which encodes both local and global spatial relationships. A Hybrid Adaptive Attention (HAA) mechanism is embedded within MSAG to selectively enhance emotion-relevant features across spatial and channel dimensions. The extracted features are subsequently passed through a Multi-Scale Simplicial Transformer (MSST), which models topological and contextual dependencies across different facial regions. Finally, a fully connected classifier uses the aggregated features to perform emotion classification. The entire flow is visually summarized in Fig. 2 for better comprehension.

3.1. Preprocessing and face detection

FER is a multi-stage process that begins with preprocessing the raw input images. Given a dataset of facial images, denoted as $X = \{x_1, x_2, \ldots, x_N\}$, where each image x_i exists in the three-dimensional space $\mathbb{R}^{(H\times W\times C)}$, the primary objective of the preprocessing stage is to extract meaningful facial regions while ensuring uniformity across all samples. The first critical step in this pipeline involves detecting and localizing faces within each image, which is efficiently accomplished using the advanced YOLOv8 object detection framework.

The YOLOv8 model partitions an input image into a grid and applies convolutional layers to predict multiple bounding boxes B_{ij} for each grid cell. Each bounding box is associated with a confidence score S_{ij}, which indicates the probability that the detected region corresponds to a face. The detection process can be mathematically formalized as follows:

$$B_{ij} = (x_{\min}, y_{\min}, x_{\max}, y_{\max}, S_{ij}) \tag{1}$$

where x_{min}, y_{min}, x_{max}, y_{max} spatial coordinates of the bounding box, and S_{ij} is the confidence score. Since multiple overlapping bounding boxes may be predicted for the same face, Non-Maximum Suppression (NMS) is applied to retain the most relevant detection. The NMS algorithm compares each pair of bounding boxes by computing the Intersection over Union (IoU), which quantifies the degree of overlap between two boxes B_{ij} and B_{ik}. It is defined as the ratio of the area of their intersection to the area of their union:

$$IoU(B_{ij}, B_{ik}) = \frac{|B_{ij} \cap B_{ik}|}{|B_{ij} \cup B_{ik}|} \tag{2}$$

Among the overlapping boxes, the one with the highest confidence score is selected as the final bounding box, ensuring minimal overlap with others based on the IoU threshold. This constraint ensures that only one bounding box is selected per detected face, with minimal overlap, based on confidence. The selected facial region is then cropped and resized to a uniform dimension of 224 × 224, providing standardized input for downstream graph-based and transformer-based processing. This step is essential for eliminating background noise and maintaining consistency across datasets.

3.2. Creation of directed graph

Following face detection, the region of interest is transformed into a structured graph $G = (V, E)$, where each node represents a facial landmark or subregion, and edges encode spatial dependencies. Nodes capture visual features extracted from patches, while directional edges represent local and non-local relationships.

To model these relationships, we utilize a Graph Attention Network (GATv2) as depicted in Fig. 2. Each node aggregates information from its neighbors using attention weights that quantify their relevance. The attention score between nodes i and j is computed using a shared transformation matrix and a non-linear activation, yielding edge weights that

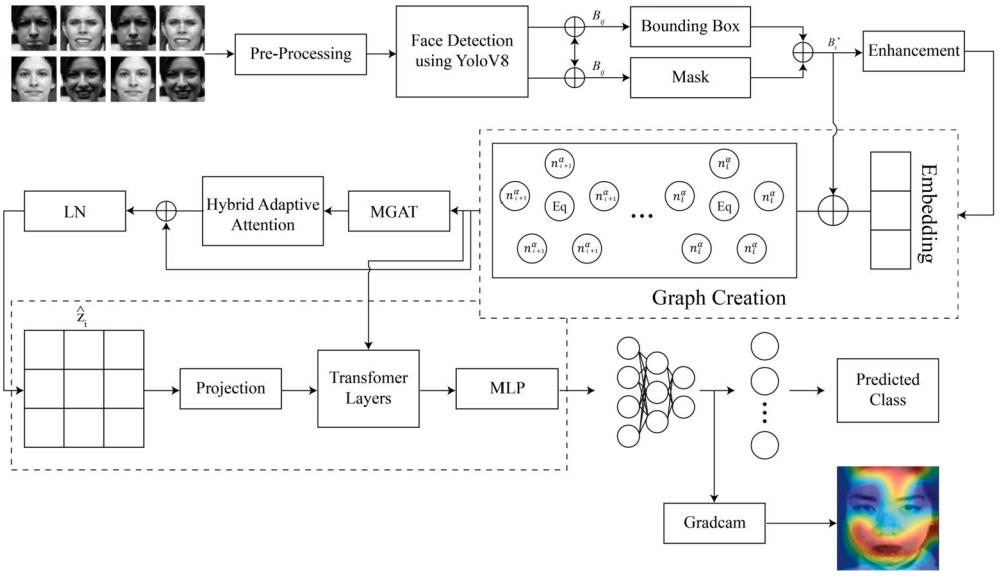

Fig. 2. Structural flow of the proposed FER system, highlighting key modules: detection, MSAG, HAA, and MSST.

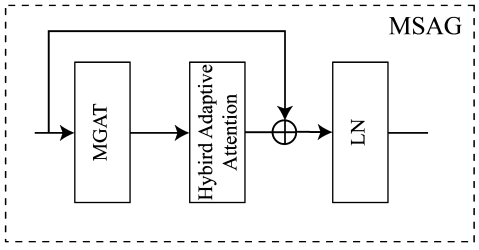

Fig. 3. Architecture of proposed MSAG module.

guide message passing. Rather than applying attention sequentially, we employ a GATv2-style update, where the joint representation of neighboring nodes is first fused, then used to compute attention:

$$\lambda_{ij} = \frac{\exp\left(a^{\top}\sigma\left(W[h_i \| h_j]\right)\right)}{\sum_{k\in\mathcal{N}(i)}\exp\left(a^{\top}\sigma\left(W[h_i \| h_k]\right)\right)} \tag{3}$$

Here, h_i is the feature vector of node i, $\mathcal{N}(i)$ is its neighborhood, W and a are learnable parameters, and σ is a non-linear function. This graph encodes short- and long-range dependencies, enabling robust spatial reasoning across facial components.

3.3. Multi-scale adaptive graph (MSAG))

Graph Neural Networks (GNNs) have an unparalleled ability to model relational data. We use three MSAGs that are the model dependency in the context of the expression in FER. In addressing the over-smoothing problem frequently encountered in GNNs [34], we create a new type of skip connection by interlinking the output of a layer with the input of the ensuing layer. That is, we use the output of a particular layer as input to the next layer, and this output is added to the input, which provides the next layer GNN with information from both the next layer and the preceding one illustrated in Fig. 3.

This particular structure of our MSAGs is intended to increase their performance by enabling more information to be maintained across layers.

$$F^{\alpha,n+1} = \mathrm{LN}\left(F^{\alpha,n} + \mathrm{MGAT}\left(F^{\alpha,n}, e^{\alpha}\right)\right) \tag{4}$$

where n ranges from 0 to N, with N representing the total number of MSAG layers. $F^{(\alpha,n)}$ denotes the feature matrix at the n-th layer, where $F^{\alpha,0} = F^{\alpha}$, and e^{α} represents the set of edges. Additionally, LN refers to the normalization layer, and MGAT denotes the multi-head graph attention layer. The MGAT mechanism processes input features using multiple attention heads, each capturing distinct relational patterns, and concatenates their outputs for a richer representation.

$$head_i^n = GAT(F^{\alpha,n}, e^{\alpha}) \tag{5}$$

The proposed framework is concerned with the aggregation of information from neighboring nodes. We can determine the relative importance of different nodes during the aggregation of patch information by calculating edge weights as specified in Eq. (3). For FER, we define our passing function as follows:

$$F_i^{\tau,n+1} = \sum_{l_j^{\tau}\in N(l_i^{\tau})} \sigma_{ij}^{\tau,n+1} P^{\tau,n+1} F_j^{\tau,l} \tag{6}$$

where $\omega_{ij}^{\alpha,n+1}$ is the attention coefficient and edge weight between node l_i^{α} and its neighboring node l_j^{α} at the n+1 layer, with i ≠ j. The feature vector $F_j^{\alpha,l}$ from the n-th layer for node l_j^{α} is included in $F^{\alpha,n}$, and $F_i^{\alpha,n+1}$ is the weight information passing at layer n+1. The parameter $F_i^{\alpha,n+1}$ is a learnable parameter. The initial input feature vector for node l_j^{α} is $F_j^{\alpha,0}$, which equals F_j^{α} with $F_j^{\alpha,0}$ being part of $F^{\alpha,0}$, where $F^{\alpha,0} = F^{\alpha}$.

3.3.1. Hybrid adaptive attention (HAA)

Spatial Attention Module: The recognition of facial expressions depends upon analyzing extracted facial features because these features do not only appear in one area but exist throughout the face to create effective mood classification. The inter-feature relationships demonstrate stability, which does not change regardless of facial shape or expression variations [35]. However, positional conditions among the variabilities of some facial landmarks depict considerable emotion realization as in Fig. 4. To improve spatial relationships among facial features, we generate a spatial attention vector that strengthens the inter-feature dependencies across the image. CBAM [36] introduced a 2D spatial attention module that uses attention computation across all channels to learn spatial feature importance. The primary purpose of CBAM was to recognize generic objects in still images, but it lacks features for learning facial element spatial relationships. The proposed facial spatial attention module adopts a concept from spatial feature representation to perform its functions as illustrated in Fig. 4. The system organizes facial characteristics into separate clusters throughout feature extraction to expand the assortment of discovered spatial patterns. The approach implements a soft-attention mechanism from [37] that engages in extracting spa-

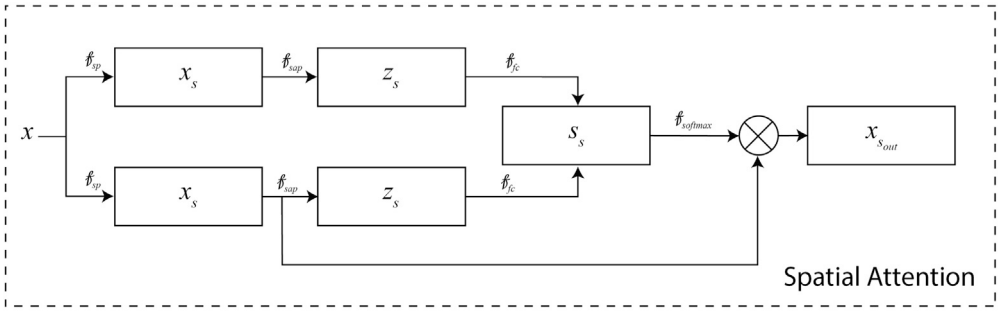

Fig. 4. Architecture of spatial attention block.

tial attention vectors with statistical feature distributions for effective global and local spatial dependency utilization versus traditional CBAM spatial attention approaches.

Soft attention enables our spatial attention mechanism to determine essential facial areas for emotion recognition measurements. The weighting system based on attention dynamics helps the system better understand features by focusing on crucial areas while reducing unimportant parts. A spatial attention vector shows identical characteristics to global spatial activation maps produced from convolutional feature extractors, leading to better facial emotion recognition through focused attention on vital facial features. The function $\mathfrak{f}_{sp}$ receives $F_j^{\tau} \in \mathbb{R}^{H\times W\times C}$ as its input feature map with dimensions H and W in space and C channels. It then produces K feature groups x_s through this composite operation.

$$x_s = \mathfrak{f}_{sp}(X) = \mathrm{ReLU}(\mathrm{BN}(\mathfrak{f}(F_j^{\tau}))) \tag{7}$$

The model contains $\mathfrak{f}_{sp}$ based on ReLU [38] and BN [39] operations. The transformed feature map $x_s \in \mathbb{R}^{H\times W\times C'\times K}$ functions as the output of a local feature extractor and uses a 3 × 3 convolution operator $\mathfrak{f}$. The number of extracted channel features is represented by C' while k represents the split groups used in this processing stage. An average pooling operation $\mathfrak{f}_{sap}$ is applied to combine extracted spatial feature groups and generate the $z_s \in \mathbb{R}^{H\times W\times C'\times K}$ spatial statistical vector.

$$z_s = \mathfrak{f}_{sap}(x_s) = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W} x_s(i,j) \tag{8}$$

We obtain the spatial attention vector by applying the fully connected operator $\mathfrak{f}_{fc}$. The practice substitutes the fully connected layer with a 1×1 convolution layer, which supports spatial consistency and decreases computational requirements. A softmax function activates the computed attention scores, which produce the final spatial attention vector s_s.

$$s_s = \mathfrak{f}_{softmax}(\mathfrak{f}_{fc}(z_s)), \quad s_s \in \mathbb{R}^{H\times W\times C'\times K} \tag{9}$$

The attention scores are normalized using the softmax function across all spatial feature groups to ensure a proper weighting distribution. The spatial attention vector receives its values through Hadamard product operations, highlighting crucial feature areas while ignoring other less essential ones. The attention-weighted spatial features $x_{s_{out}}$ result from the following computation:

$$x_{s_{out}} = s_s \odot x_s \tag{10}$$

Hadamard Product operation (⊙) applies attention values to their specific feature positions in the input data. This operation produces $x_{s_{\mathrm{out}}} \in \mathbb{R}^{H\times W\times(C'\times K)}$ as a 3D tensor and the final channel representation has dimension $C'\times K$ with spatial features grouped.

Channel Attention: For FER and similar computer vision operations, convolutional networks apply equal importance to all feature channels, which produces redundant information with diminished signal value. The recognition of expressions depends differently on various channels since some channels distribute key discrimination features, while other channels detect secondary or weak information. We have utilized a channel attention mechanism to dynamically reweigh feature channels so that the model can enhance significant features and suppress irrelevant ones. The network's decision-making process becomes focused on emotion-related areas through this mechanism, which leads to better feature representation while increasing resistance to interference and improving performance on various datasets. Fig. 5 depicts a channel attention procedure that improves feature representation quality through dynamic changes in channel significance.

The input feature map $F_M \in \mathbb{R}^{H\times W\times C}$ contains spatial dimensions H and W together with C channels. The channel significance calculation uses two global pooling methods, global max pooling, and global average pooling, to extract diverse feature elements. The global max pooling method selects the most salient features across all spatial positions and defines the operation as:

$$C_{max} = \max F_M(i,j,c), \quad \forall c, \quad C_{max} \in \mathbb{R}^{1\times1\times C} \tag{11}$$

Global average pooling maintains the complete contextual information by calculating mean channel activations.

$$C_{avg} = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W} F_M(i,j,c), \quad \forall c, \quad C_{avg} \in \mathbb{R}^{1\times1\times C} \tag{12}$$

The combined feature vectors are passed to Multi-Layer Perceptron (MLP), which includes two fully connected layers with ReLU and sigmoid activation in between. The final channel attention vector A_C is achieved by combining max-pooled and average-pooled scores through element-wise summation.

$$A_C = \sigma((W_2\delta(W_1 C_{max}) + (W_2\delta(W_1 C_{avg})) \tag{13}$$

W_1 and W_2 are trainable weight matrices in the expression, while ReLU activation is denoted by $\delta(\cdot)$ and $\sigma(\cdot)$ represents sigmoid functions to keep [0,1] output range.

The attention weight gets applied by Hadamard multiplication to the input feature map and generates the recalibrated feature map.

$$Att = A_C \odot F_M \tag{14}$$

In HAA, the sequence integrates spatial and channel attention to enhance the feature representation of the input data. Initially, spatial attention is applied through the operation $x_{s_{out}} = s_s \odot x_s$, where s is the spatial attention weights and x_s is the spatial component of the input. This focuses the model on important spatial features by selectively emphasizing specific input regions. The channel attention component A_C modifies the feature map F_M, effectively recalibrating the channel-wise features through the operation $Att = A_C \odot F_M$. By sequentially applying this attention, the model can leverage the critical spatial locations and the most informative feature channels. The final output x_{out} is produced by combining these two processed components, represented by $x_{out} = (A_C \odot F_M) \odot (s_s \odot x_s)$.

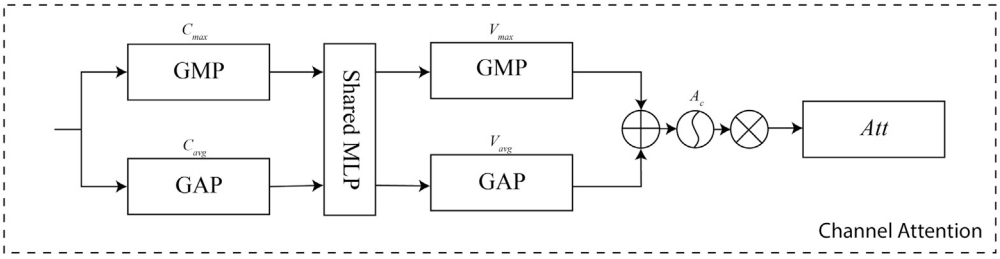

Fig. 5. Architecture of channel attention block.

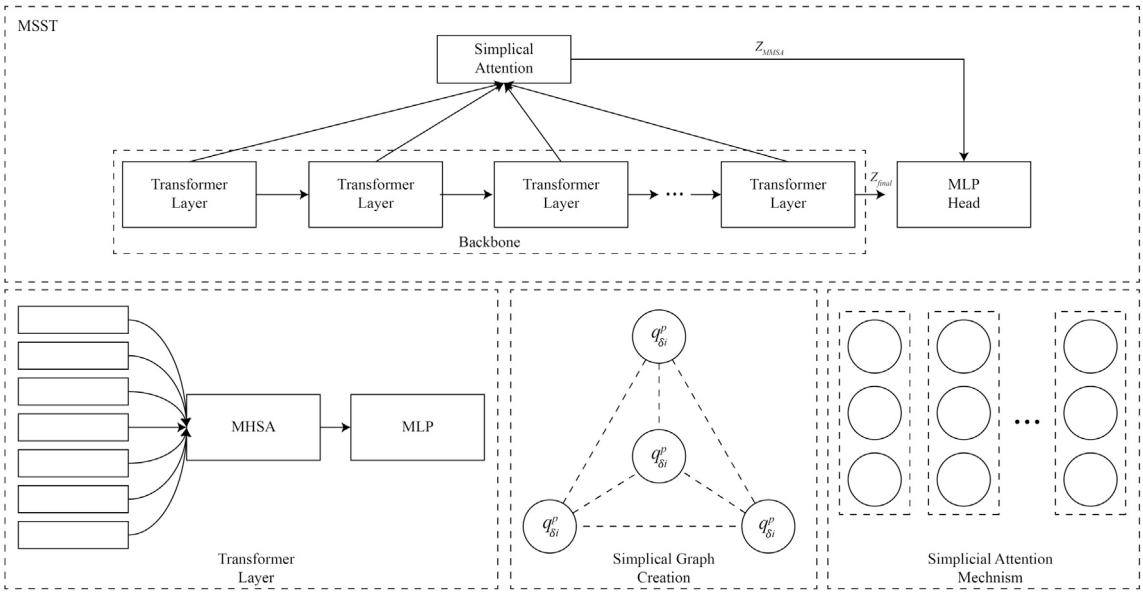

Fig. 6. Architecture of proposed MSST module.

3.4. Multi-scale simplicial transformer (MSST)

The MSST enhances feature representation through Transformer-based hierarchical processing of information and multi-scale dependencies. MSST incorporates MSSA, which allows the transfer of structured data between Transformer layers to obtain local and global relationship-based features through its designed structure. Fig. 6 employs Transformer layers through which MSSA obtains and refines core structural elements at multiple scales.

3.4.1. MSST backbone

The input features $x_{out} \in \mathbb{R}^{H\times W\times C}$ gets divided into fixed-length non-overlapping patches which form the set $x_p \in \mathbb{R}^{N\times(P_2\times C)}$. The patch size is represented by P while the total number of patches equals $N = HW/P_2$. In our approach, we set P =16. The processed image patches form a sequence that can be handled by a standard transformer model directly. Before feeding the first transformer layer, the flattened patches undergo a linear projection to establish a uniform dimension D so that the initial input becomes $x_1 \in \mathbb{R}^{N\times D}$. As described in Vision Transformer (ViT) [40], we embed trainable positions alongside a classification token to maintain spatial relations in input images. Using a classification token the final input becomes $x_1 \in \mathbb{R}^{(N+1)\times D}$.

After processing the image input through the MSAG block, MSST architecture presents data flow in two primary components: the MSST backbone and the MSSA. The MSST backbone produces deep feature output Z_{out} by passing through multiple transformer layers, as shown in Fig. 6. The basic unit of operations within a transformer layer combines Multi-Head Self-Attention (MHSA) operations with MLP computations. Both components' MHSA and MLP blocks adopt residual connections with LN implemented before each block operation. MHSA processes information at various points with multiple heads. For the input tensor $x_i \in \mathbb{R}^{(N+1)\times D}$ of the i-th layer, the output of the MHSA block is represented as

$$\hat{z}_i = x_i + MHSA(x_i) \tag{15}$$

The MLP block comprises two fully connected layers. An activation function, such as GELU [41], is situated between them. The output z_i in the i-th layer is expressed as:

$$z_i = \hat{z}_i + MLP(\hat{z}_i) \tag{16}$$

The output z_i in the i-th layer serves as the input x_{i+1} for the (i + 1)-th layer. The output of the deep feature corresponds to the output of the final transformer layer.

$$Z_{final} = z_l \tag{17}$$

The MSSA receives embedded features from numerous Transformer processing stages. The embeddings of an input image divided into patches enter a series of Transformer layers. MSSA performs its operations on representations that come from MHSA and an MLP through each processing layer. Through the MSSA module, the selected layers in the network provide feature representations for analyzing multi-scale interactions. The module functions by receiving outputs from various transformer layers and utilizes a simplicial graph attention mechanism to find meaningful structural relationships across different components.

3.4.2. Multi-scale simplicial attention (MSSA)

MSSA adds a module to Transformer architecture to achieve multi-level feature extraction across transformer layers. MSSA block obtains essential feature patterns from different transformer layers and enables

multifaceted information exchange between these layers. The feature representation updates of a given simplex run through message passing operations performed with its adjacent simplices across each transformer layer. The p-simplicial δ_i^p, δ_j^p attention relationship between two elements is evaluated through the calculation of:

$$\tau_{ij}^p = \text{LeakyReLU}\left(\gamma_p^T \left[M_p q_{\delta_i}^p \parallel M_p q_{\delta_j}^p \parallel M_{p+1} q_{\delta_{ij}}^{p+1}\right]\right) \tag{18}$$

Where $\parallel$ signifies concatenation, γ_p represents a learnable attention vector, and M_p denotes the weight matrix for a linear transformation that is specific to the p-order simplices and shared among all simplices of that order. The feature vector $q_{\delta_{ij}}^{p+1}$ belongs to the p+1 simplex that contains both δ_i^p and δ_j^p as faces. To obtain the normalized attention scores, the softmax function is applied to the unnormalized attention values as follows:

$$\omega_{ij}^p = \frac{\exp(\tau_{ij}^p)}{\sum_{r\in N(\delta_i^p)} \exp(\tau_{ir}^p)} \tag{19}$$

where $N(\delta_i^p)$ denotes the set of neighboring simplices of δ_i^p. The learned weight coefficients are then used to aggregate feature information from the neighboring simplices, leading to an updated feature representation for each p simplex, which is computed as:

$$q_{\delta_i}^p = \sigma\left(\sum_{j\in N(\delta_i^p)} \omega_{ij}^p M_p q_{\delta_j}^p\right) \tag{20}$$

where $\sigma(\cdot)$ denotes a nonlinear activation function. Multi-head attention is incorporated by employing R independent attention mechanisms to improve training stability. The outputs from these mechanisms are concatenated to generate the final learned feature representation for each simplex:

$$Z_{MSSA} = \sum_{r=1}^{R} \sigma\left(\sum_{j\in N(\delta_i^p)} \omega_{ij}^{(p,r)} M_p^r q_{\delta_j}^p\right) \tag{21}$$

Where $\omega_{ij}^{p,r}$ and M_p^r denote the attention weights and transformation matrices associated with the r-th attention head, respectively. The final output for the p-order simplices, denoted as S_{λ_p}, consists of the learned embeddings for all simplices at that order: $S_{\lambda_p} = \{q_{\delta_1}^p, q_{\delta_2}^p, \ldots, q_{\delta_{|\lambda_p|}}^p\}$. Since the model employs P graph attention layers, it produces P sets of simplicial complex-specific embeddings: $S_{\lambda_0}, S_{\lambda_1}, \cdots, S_{\lambda_{P-1}}$. Depending on the downstream task, a subset of these embeddings is utilized. For example, in node classification tasks, only the embeddings corresponding to S_{λ_0} are used in the final softmax layer. By integrating the deep feature output Z_{final} (Eq. (17)) with the simplicial feature output (Eq. (21)), the final output Z_{out} is formulated as:

$$Z_{out} = Z_{final} + \lambda \cdot Z_{MSSA} \tag{22}$$

Where λ is a scaling factor that adjusts the relative contribution of the two feature types. The MSST backbone leverages multiple transformer layers to transform image input into a more interpretable feature representation for the model. Through MSSA modules, the model gains improved capabilities to simultaneously process multi-scale features from different transformer blocks. The MSSA system gets its information from various layers to build feature representation with greater depth. The model uses this method to concentrate on vital information by discarding unnecessary details.

As it operates inside the MSSA block, the simplicial connection enhances the spread of attention between layers, leading to optimized communication between information streams. The MSSA block helps MSST focus on significant features better to deliver high classification accuracy. The absence of additional learnable parameters in MSST enables it to be integrated without complexity growth when applied to attention-based models.

4. Experimental results

This section studies the proposed facial emotion recognition model's experimental results, along with details about datasets and implementation setup, ablation analysis, and a comparison with state-of-the-art approaches. Different evaluation metrics measure the proposed model's performance.

4.1. Datasets

The proposed model is tested on four publicly available datasets, including AffectNet FER2013, CK+, Occlusion, and Pose AffectNet. AffectNet [9] is the most enormous uncontrolled facial expression dataset since it contains 450,000 manually labeled images collected from multiple online search engines. The dataset raises many issues because it includes diverse ethnicities, numerous poses, various obstacles, and lighting situations, different backgrounds, multiple complex elements, and an unbalanced distribution of categories. The FER2013 [8] provides static image data instead of video sequences, distinguishing it from other datasets. As the dataset consists of 35,886 facial expression images that measure 48×48 pixels, the collection features seven distinct facial expressions. The data distribution consists of 28,708 images for training and 3,589 images for validation and testing. The research employs the FER2013's original training and test databases for its model development process.

The CK+ [5] dataset extends the Cohn-Kanade (CK) dataset through 593 video sequences and 123 subjects with annotation of 327 sequences demonstrating seven basic expressions as well as contempt. The expressed emotions begin at a non-emotional stage in the initial frames before reaching the maximal expression in the concluding frames. Previous research points to the first image as showing a neutral face, while the last image displays the desired emotional state. A total of 618 images were labeled with seven basic emotions, while 654 images received labeling for the same eight emotions for validation across datasets. Occlusion and pose variation datasets [14] consists of six test sets for pose variation and occlusion, including RAF-DB, FERPlus, and AffectNet-8. This study utilizes AffectNet Occlusion and Pose Variation datasets. There are 683 samples within the occlusion subset. The pose subset contains samples above 30 and 45 degrees. The pose-AffectNet subset includes two groups of samples containing 1949 and 985 images reflecting various position angles.

4.2. Implementation details and evaluation metrics

The facial emotion recognition model utilizes PyTorch, TensorFlow, and Keras alongside transformers to perform computation. At the same time, an NVIDIA RTX 4090 GPU works for hardware acceleration to achieve quick computation and faster convergence during training. The implemented architecture utilizes batch normalization alongside the Adam optimizer at a 0.2 dropout rate, which assists in overfitting prevention and improves generalization performance. The processed data runs on a batch size of 8, while the data distribution includes training (70%), testing (20%), and validation (10%). The learning rate decreases automatically through a scheduler when the validation error grows, while early stopping maintains training by stopping at epoch 10 if there is no performance improvement.

To evaluate the proposed approach fairly, a variety of standard and statistically rigorous performance metrics were used. Accuracy (Acc) measures the percentage of correctly identified samples, however class imbalance may influence it. The Macro F1-score (F1) calculates the unweighted mean of per-class F1-scores to treat each category equally regardless of frequency. Each class is also assessed using Precision (Pre) and Sensitivity (Sen) metrics. Precision measures projected label correctness, whereas sensitivity represents the model's ability to find all relevant instances. To improve evaluation in multi-class settings, Cohen's Kappa coefficient (Kappa) is used to estimate inter-rater agreement

Table 1
Performance comparison of the proposed model on AffectNet and FER2013 datasets.

Class	AffectNet				FER2013			
	Pre (%)	Sen (%)	F1 (%)	Kappa (%)	Pre (%)	Sen (%)	F1 (%)	Kappa (%)
Anger	79.75	79.75	79.75	76.86	90.20	91.23	90.71	89.86
Contempt	80.30	81.50	80.89	78.81	-	-	-	-
Disgust	80.29	83.50	81.86	81.03	63.45	82.88	71.88	82.53
Fear	81.57	83.00	82.28	80.52	87.39	90.72	89.03	89.11
Happy	77.01	81.25	79.08	78.40	94.92	88.50	91.60	85.06
Neutral	85.86	85.00	85.43	82.88	89.91	93.27	91.56	91.81
Sad	84.45	78.75	81.50	75.95	92.28	89.09	90.66	86.90
Surprise	86.32	82.00	84.10	79.57	88.31	91.82	90.03	90.70
Accuracy	81.84				90.40			
Macro Average	81.94	81.84	81.86	79.25	86.64	89.65	87.92	87.99
Weighted Average	81.94	81.84	81.86	79.25	90.65	90.40	90.46	87.99

adjusted for chance, providing more precise results. In uneven class distribution circumstances, the Matthews Correlation Coefficient (MCC) evaluates classification performance more fairly than accuracy. Using Spearman's Rank Correlation coefficient (ρ), we analyze the monotonic relationship between expected and actual label distributions. Multiple sample alignment of expected and actual class ranks is assessed by this statistic. The paired t-test and Wilcoxon signed-rank test are used to examine the statistical significance of performance improvements over baseline techniques from multiple independent runs. Testing shows that the proposed method's superiority across datasets is not attributable to random fluctuation.

4.3. Model performance

The performance evaluation of the model for the AffectNet and FER2013 datasets appears in Table 1 to demonstrate its effectiveness for generalizing expressions in faces. This classification method reaches an 81.84% accuracy rate, indicating successful emotion differentiation between different categories. The model demonstrates uniform recognition abilities across every category because the weighted and macro F1-score measure reaches 81.86%. The model reliably processed real-world facial expression variations because its Cohen's Kappa score reaches 79.25%. Certain expressions show variations in score distribution, most probably due to the combination of facial subtleties and unbalanced data distribution in the dataset. The model performance analysis on the FER2013 dataset demonstrates 90.40% success in identifying various emotional expressions. The model displays balanced class performance through F1-score values of 87.92% on the macro scale and 90.46% on weighted average data. The predictions match actual labels with high agreement according to Cohen's Kappa score of 87.99%. The model stands out in identifying happy, neutral, and angry emotions because it reaches more than 90% F1-scores, demonstrating a clear distinction among these expressions. Disgust represents a challenge to the model since its F1-score reaches only 71.88%, although this indicator reflects how accurately the model identifies this emotion class. Although the model experiences less successful results while identifying disgust emotion, its high recall results across other categories ensure essential expressions are appropriately captured. It is important to note that the performance values visually represented in confusion matrix in Fig. 7 are quantitatively derived from the classification statistics shown in Table 1. However confusion matrix in Fig. 7 provides more detailed analysis showcasing number of correct and incorrect prediction along each class.

Fig. 8 presents the training and validation accuracy and loss curves for AffectNet and FER2013 in the bottom row. In Fig. 8(a), the model shows a steady increase in accuracy with minimal difference between training and validation, indicating strong generalization. The loss decreases smoothly, suggesting effective learning with no signs of overfitting. In Fig. 8(b), the model demonstrates an even better learning trend, with higher accuracy and lower loss. The validation curve closely follows the training curve, confirming substantial feature extraction and generalization. The results indicate that the model performs well on both datasets, showing faster convergence and better learning stability on FER2013.

4.4. Statistical evaluation of model performance

The model performance assessment on AffectNet and FER2013 occurs through Table 2 using Matthews Correlation Coefficient (MCC) along with Spearman's Rank Correlation (ρ) which provide metrics that go beyond standard accuracy scores. MCC is highly suitable for detecting class biases in datasets because it rates accuracy by evaluating relationships between actual positive outcomes and false detections alongside correct negative and false negatives. The model demonstrates a reliable calibration and unbiased prediction capability when its MCC score reaches high values. Spearman's Rank Correlation evaluates the rank preservation between model predicted and actual emotion ordering to ensure correct ranking dependencies between expressions in predicted outputs. Aside from accuracy assessment, these tests assess model-predicted classifications and consistency and ranking accuracy for actual emotional intensity variations in real-world applications. The model demonstrated reliable performance by achieving MCC scores of 0.79 and 0.88 and Spearman correlation values of 0.80 and 0.88 on AffectNet and FER2013, respectively. This demonstrated that it generalized effectively on FER2013 data. The model shows its effectiveness in maintaining class consistency and accurately identifying emotional expression strengths, making it work optimally for complex FER applications.

In Table 3, the paired t-test and Wilcoxon signed-rank test were employed to determine if the performance differences between the proposed model and competing baselines are statistically significant. The Wilcoxon test is a strong non-parametric alternative that does not imply normal distribution, but the paired t-test was chosen for its sensitivity to mean differences under normal distribution, making the analysis more reliable across several performance distributions. Every model was trained and evaluated five times separately, and accuracy, F1-score, and Cohen's Kappa were compared statistically. AffectNet, FER2013, and CK+ benchmark datasets and three representative baseline models—LAENet-SA (transformer-based), FGENet (graph-based), and VTFF (CNN-based)—were systematically tested. The null hypothesis always predicts no noticeable performance difference between the proposed model and the matched baseline, while the alternative hypothesis proposes a performance advantage for the technique. All t-test and Wilcoxon test p-values were below 0.05 across all comparisons, datasets, and measurements. The model predicted a statistically significant increase in F1-Score on the AffectNet dataset compared to LAENet-SA, with a Wilcoxon p-value of 0.007 and a t-test p-value of 0.003. Similar relevance was found for accuracy and kappa across the other datasets, confirming the notion that performance increases reflect meaningful and

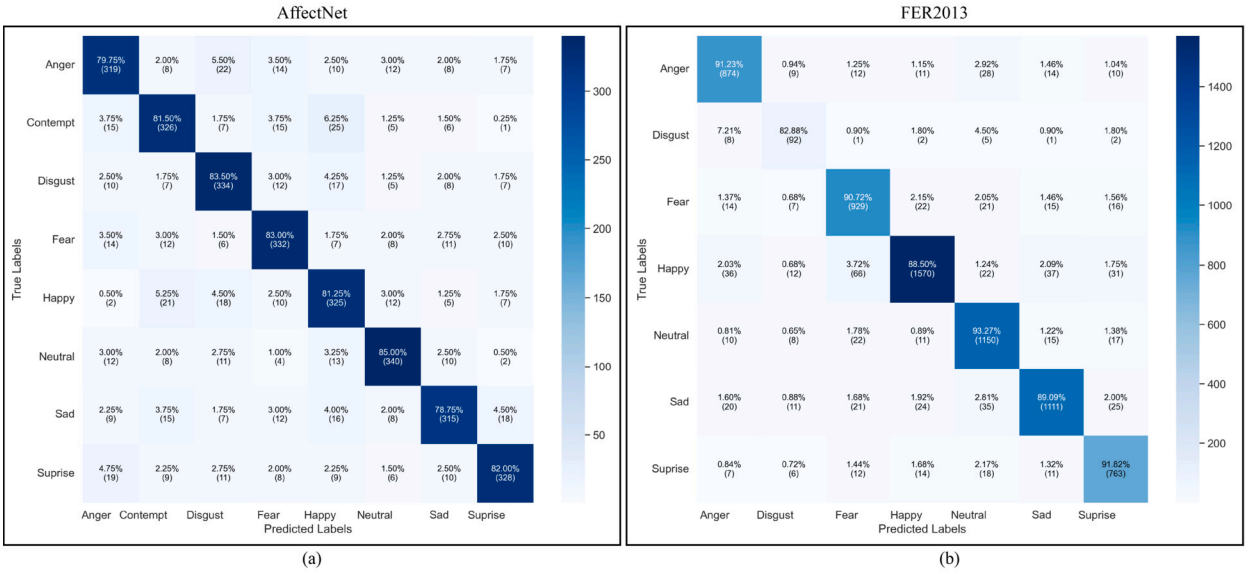

Fig. 7. Confusion matrix for AffectNet and FER2013 dataset.

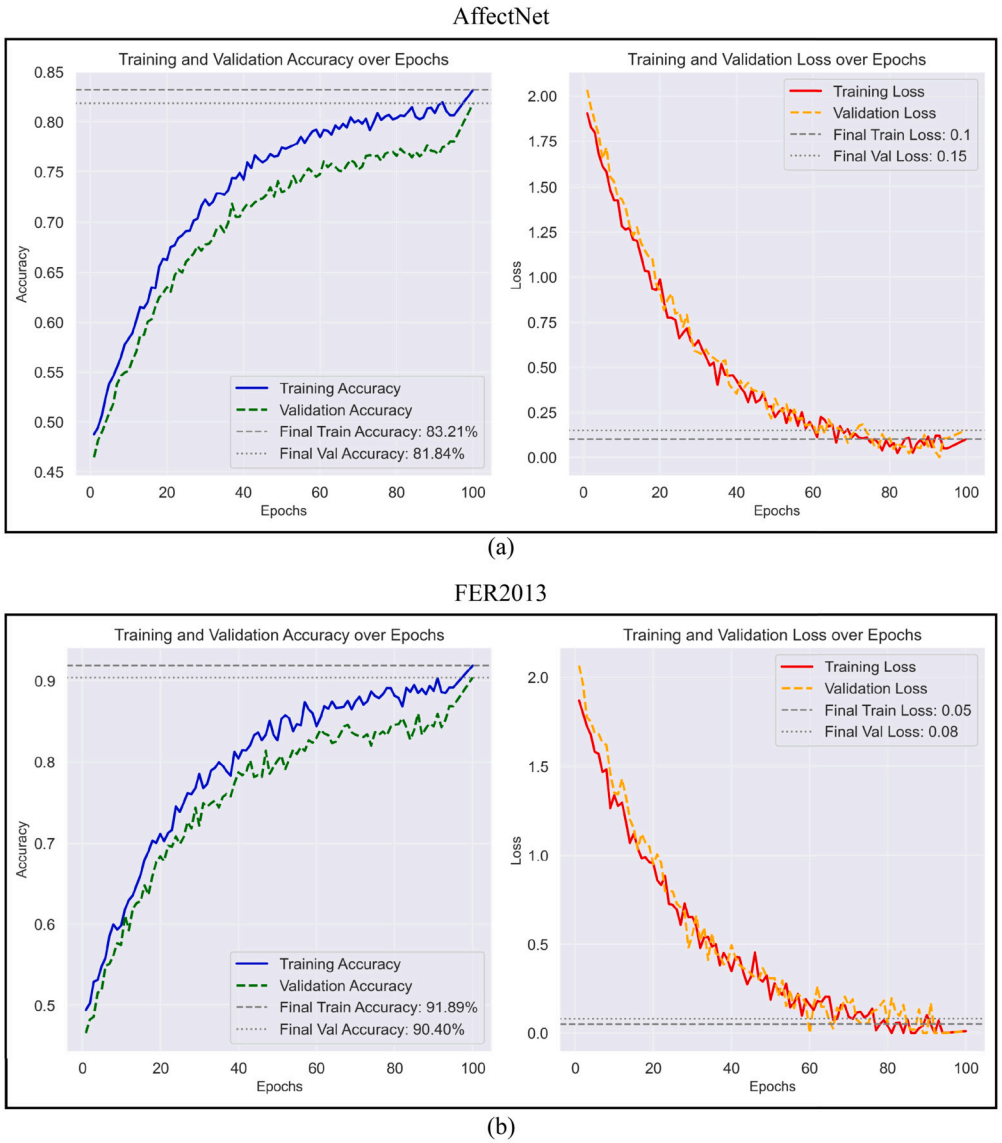

Fig. 8. Training and validation accuracy and loss curves for AffectNet and FER2013 dataset.

Table 2
Statistical results of the proposed model on AffectNet and FER2013 datasets.

Dataset	Matthews Correlation Coefficient (MCC)	Spearman's Rank Correlation (ρ)
AffectNet	0.79	0.80
FER2013	0.88	0.88

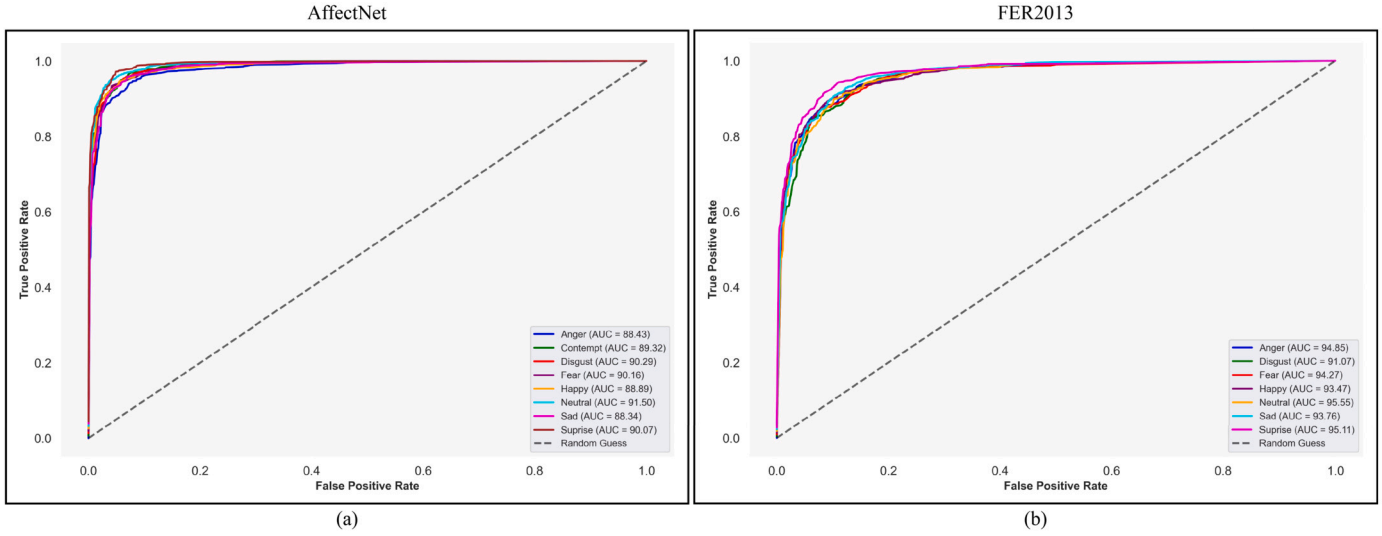

Fig. 9. ROC curves of the proposed model on AffectNet and FER2013 datasets.

Table 3
Statistical significance testing results using Paired t-test and Wilcoxon Signed-Rank Test.

Dataset	Metric	Baseline	t-test p-value	Wilcoxon p-value
AffectNet	Accuracy	LAENet-SA	0.014	0.021
	F1-Score	LAENet-SA	0.003	0.007
	Kappa	LAENet-SA	0.009	0.012
FER2013	Accuracy	FGENet	0.019	0.024
	F1-Score	FGENet	0.008	0.015
	Kappa	FGENet	0.005	0.009
CK+	Accuracy	VTFF	0.041	0.038
	F1-Score	VTFF	0.027	0.035
	Kappa	VTFF	0.030	0.031

sustained improvements over current techniques rather than random variation.

4.5. ROC curve analysis

Fig. 9 demonstrates how the model performs classification tasks across two datasets through its AUC values for each facial expression. The model in Fig. 9(a) demonstrates high AUC measurement success across different emotions on the AffectNet dataset, yet the anger and sadness classes display marginal separability compared to others. The classification performance from Fig. 9(b) demonstrates better generalization throughout the facial expressions, showing elevated AUC values across all expressions. Surprise, neutral, and sad expressions display outstanding recognition accuracy in the model's assessment of the FER2013 dataset because it exhibits enhanced capabilities of detecting features specific to facial expressions.

4.6. Cross dataset validation on CK+

A cross-dataset performance evaluation of different training datasets against CK+ is provided in Table 4. AffectNet training data provides significantly superior generalization capabilities compared to FER2013 training data because previous models [25] and [31] reach 90.80% and 58.10% accuracy, respectively, thus showing that FER2013 might not have enough diverse or high-quality labels for robust feature extraction. The research yielded superior results when using this model as it reached 97.89% AffectNet accuracy and 93.67% FER2013 accuracy because it extracted potent emotion-relevant features that were easily transferable. The best accuracy level of 98.54% emerges when the training data includes a mix of AffectNet and FER2013 because this multi-domain approach boosts feature variety while reducing dataset biases and enhancing performance on unknown datasets.

4.7. Performance on occlusion and pose variation dataset

Table 5 compares model performance through examinations of the results on Occlusion tests together with Pose-AffectNet data at two head position angles of 30° and 45°. According to results, AMP-Net [24] reached 64.27% accuracy during occlusion testing, while MRAN [44] demonstrated higher success with 62.05% accuracy in a 45° head position. These methods successfully detect minor occlusions and pose changes. GFFT [42] and HALNeT [43] demonstrate comparable performance patterns to eachother even though these methods bring attention mechanisms or feature aggregation functionalities to the system but fail at handling significant fluctuations. The proposed model displays superior performance compared to all existing models by reaching 77.18% on Occlusion-AffectNet, 75.79% on Pose 30°, and 75.45% on Pose 45°, which reflects its robustness when handling occlusion and pose variations.

4.8. Ablation studies

4.8.1. Evaluation of components

The performance of YOLOv8 (Baseline) and the enhanced models that use MSAG and MSST run on AffectNet and FER2013 datasets appears in Fig. 10. The baseline YOLOv8 model starts at 62.42% AffectNet accuracy and 59.81% FER2013 accuracy, indicating its essential ability to conduct FER. The combination of MSAG using HAA with spatial and channel techniques results in improved accuracy levels of 74.79%

Table 4
Generalization of proposed model on CK+ dataset.

Model	Train	Test	Acc (%)
Light Attention Embedding Network (LAENet-SA) [25]	AffectNet	CK+	90.80
Efficient-SwishNe (SwishNet) [31]	FER2013	CK+	58.10
Proposed	AffectNet	CK+	97.89
	FER2013	CK+	93.67
	AffectNet+FER2013	CK+	98.54

Table 5
Performance comparison of different methods under occlusion and pose AffectNet.

Method	Occlusion	Pose 30°	Pose 45°
Adaptive Multilayer Perceptual Attention Network (AMP-Net) [24]	64.27	61.37	61.16
Visual Transformers with Feature Fusion (VTFF) [44]	62.98	60.61	61.00
Global-local feature fusion (GFFT) [42]	65.04	61.49	61.35
Fine-Grained Associative Graph Representation (FG-AGR) [45]	64.24	61.26	61.15
Multi-Relations Aware Network (MRAN) [46]	63.54	61.42	62.05
Hybrid Attention-Aware Learning Network (HALNeT) [43]	63.54	61.52	61.32
Proposed	**77.18**	**75.79**	**75.45**

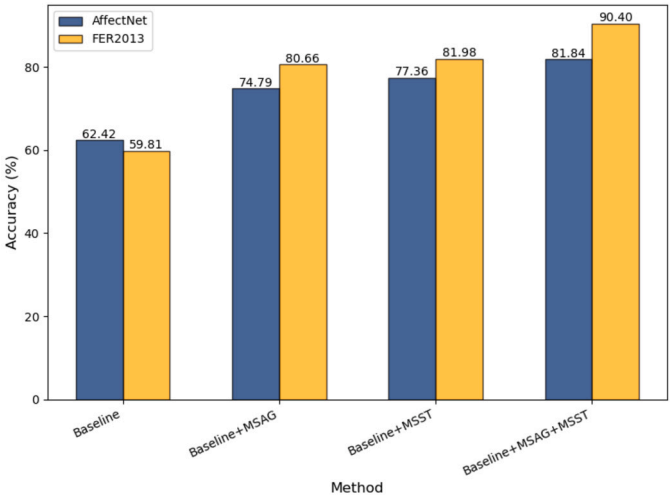

Fig. 10. Component wise comparison of proposed model on AffectNet and FER2013 datasets.

on AffectNet and 80.66% on FER2013 by effectively detecting complex face relationships. The implementation of MSST that adopts MSSA for spatial-temporal feature propagation resulted in 77.36% accuracy on AffectNet and 81.98% accuracy on FER2013 due to its effective multi-scale feature learning capabilities. Hybrid reuse of MSST and MSAG results in the best training performance at 81.84% accuracy on AffectNet and 90.40% accuracy on FER2013, as these attention mechanisms support each other for improved FER results.

4.8.2. Feature attention visualization

Grad-CAM enables the visual identification of significant facial areas, which assist emotion recognition tasks in Fig. 11. The model focuses its analysis on specific areas of the facial region that appear in different heat layers based on their priority levels during classification. The model primarily concentrates on viewing the eyes alongside the mouth and forehead since these facial characteristics are essential for emotion identification. The model directs its attention firmly toward the mouth region when detecting positive expressions, while negative expressions lead to increased attention toward brows and lips. Such visual insight allows observation of the decision mechanism used by the model to guarantee that its processing methods match human emotional interpretation.

4.8.3. Feature distribution visualization

Subsequent evaluation of our methodology includes t-SNE-based visualization for the 2D feature mappings learned by the proposed model across two different datasets. Fig. 12(a) shows feature embeddings from the AffectNet dataset, which display compact clusters between instances and distinct class separations, thus proving the model successfully detects emotional features. The feature distributions of the FER2013 dataset presented in Fig. 12(b) formed distinct clusters, yet these clusters showed minor distribution changes because of specific characteristics within the dataset. The visualizations validate that the proposed method produces distinct and compact features within classes while separating emotions into distinct regions, ultimately resulting in generalized emotion recognition across different datasets.

4.9. Comparison with existing models

The comparison in Table 6 demonstrates the outcome of facial expression recognition techniques against AffectNet and FER2013 data, which highlights the best approaches for dealing with real-world conditions. AffectNet challenges models CF-DAN [29] and HLA-ViT [30] because of their diverse and uncontrolled expressions, even though both achieve accuracy rates at 63.58% and 65.07%. The proposed method achieves superior feature representation and generalization capabilities, surpassing the other techniques by 81.84%. Circumstances in the FER2013 framework create difficulties for models due to high intra-class variations, but the proposed method achieves 90.40% while AMP-Net [24] and FGENet [27] obtain 74.48% and 70.49%. The approach benefits from its advanced spatial attention and feature fusion integration, thus making it more effective at detecting difficult and complex emotional expressions.

4.10. Practical implications and deployment considerations

Facial emotion recognition systems must be effectively implemented, balancing model accuracy, computational efficiency, and reliability across various circumstances. The proposed architecture is designed for practical applications, particularly those necessitating advanced understanding of human emotions in uncontrolled environments. Adaptive educational systems that respond to student emotions, mental health surveillance for the early detection of stress or depression, and human-computer interface (HCI) platforms that empathetically respond to user expressions are potential applications.

Despite its complexity, the model is compatible with contemporary high-performance computing systems that utilize GPUs often located in institutional servers and cloud-based APIs. For desktop or kiosk-based

Fig. 11. Grad-CAM visualizations highlighting key facial regions for emotion recognition.

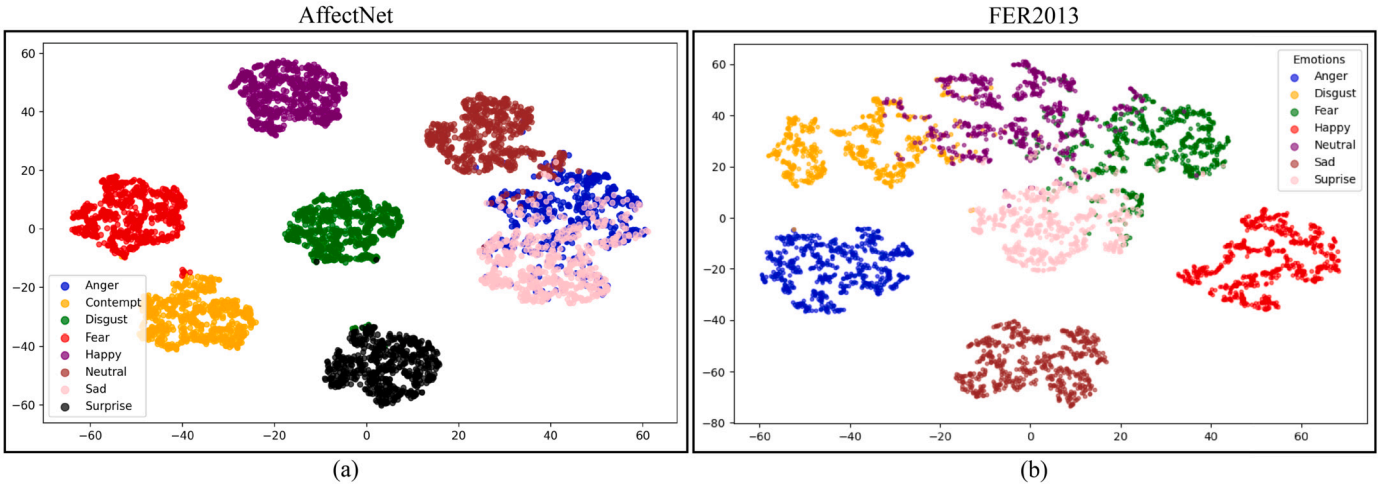

Fig. 12. t-SNE visualization of feature distributions learned by proposed approach on AffectNet and FER2013 datasets.

Table 6
Comparison with existing FER methods on AffectNet and FER2013.

Method	Datasets	Acc (%)
Relative Uncertainty Learning (RUL) [21]	AffectNet	60.66
Identity and Pose Disentangled (IPD-FER) [23]		62.23
Visual Transformers and Feature Fusion (VTFF) [28]		61.85
Cross-fusion dual-attention network (CF-DAN) [29]		63.58
Hybrid Local Attention-based Transformer (HLA-ViT) [30]		65.07
Proposed		**81.84**
Relative Uncertainty Learning (RUL) [21]	FER2013	73.75
Point Adversarial Self Mining (PASM) [22]		73.59
Adaptive Multilayer Perceptual Attention Network (AMP-Net) [24]		74.48
Three stage feature selection [26]		66.10
Facial Geometry Enhanced Network (FGENet) [27]		70.49
Proposed		**90.40**

systems, the inference time per image is approximately 20–30 ms, which is suitable for near real-time applications. To reduce memory and computational overhead in mobile or edge computing contexts, the current approach may necessitate pruning or distillation approaches. These enhancements could facilitate implementation on devices such as embedded vision systems utilized in retail or automotive settings, as well as smartphones and tablets.

5. Conclusion

The research presents an innovative FER framework that successfully tackles problems involving occlusions, pose variations, and subtle expression variations. Using graph-based spatial reasoning alongside multi-scale feature extraction produces better and more adaptable characteristic representations than standard methods. The proposed model delivers superior performance in FER2013 and AffectNet evaluations by attaining accuracy rates of 81.84% and 90.40%, which outperforms existing state-of-the-art approaches. This model shows consistent performance during cross-dataset validation tests on CK+, which demonstrates strong generalization abilities because it reaches 97.89% accuracy when trained on AffectNet and 93.67% on FER2013 while reaching 98.54% accuracy with combined training from both datasets. Several performance constraints exist in this approach because allocating significant computing power slows real-time system execution. Future work will direct efforts toward simplifying the model while developing self-supervised learning techniques and domain-generalization methods to boost system responsiveness across various application scenarios. The proposed framework enhances affective computing and human-computer interaction alongside psychological analysis by providing an improved FER solution that is both dependable and scalable.

CRediT authorship contribution statement

Samia Nawaz Yousafzai: Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Conceptualization. **Inzamam Mashood Nasir:** Writing – original draft, Validation,

Resources, Investigation, Data curation, Conceptualization. **Oumaima Saidani:** Writing – review & editing, Validation, Resources, Project administration, Investigation, Conceptualization. **Refka Ghodhbani:** Writing – review & editing, Validation, Resources, Project administration, Investigation, Conceptualization. **Yeonghyeon Gu:** Writing – review & editing, Supervision, Methodology, Funding acquisition, Conceptualization. **Muhammad Syafrudin:** Writing – review & editing, Visualization, Supervision, Methodology, Funding acquisition, Formal analysis, Conceptualization. **Norma Latif Fitriyani:** Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation.

Funding

This research was supported by the IITP (Institute of Information & Communications Technology Planning & Evaluation)–ITRC (Information Technology Research Center) grant funded by the Korea government (Ministry of Science and ICT) (IITP-2025-RS-2024-00437191). Princess Nourah Bint Abdulrahman University Researchers Supporting Project number (PNURSP2025R760), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was supported by the IITP (Institute of Information & Communications Technology Planning & Evaluation)–ITRC (Information Technology Research Center) grant funded by the Korea government (Ministry of Science and ICT) (IITP-2025-RS-2024-00437191). The authors extend their appreciation to the Deanship of Scientific Research at Northern Border University, Arar, KSA for funding this research work through the project number “NBU-FPEJ-2025-2461-06”. Princess Nourah Bint Abdulrahman University Researchers Supporting Project number (PNURSP2025R760), Princess Nourah Bint Abdulrahman University, Riyadh, Saudi Arabia.

Data availability

This research has been conducted on four publicly available datasets including CK+ [5], FER2013 [8], AffectNet [9] and Occlusion and Pose Dataset [14]. The source code is available at https://github.com/Samia-Nawaz/Facial-Emotion-Recognition.

References

[1] Liu Z, Wu M, Cao W, Chen L, Xu J, Zhang R, et al. A facial expression emotion recognition based human-robot interaction system. IEEE/CAA J Autom Sin 2017;4(4):668–76.
[2] Fei Z, Yang E, Li DD-U, Butler S, Ijomah W, Li X, et al. Deep convolution network based emotion analysis towards mental health care. Neurocomputing 2020;388:212–27.
[3] Bisogni C, Castiglione A, Hossain S, Narducci F, Umer S. Impact of deep learning approaches on facial expression recognition in healthcare industries. IEEE Trans Ind Inform 2022;18(8):5619–27.
[4] Pantic M, Valstar M, Rademaker R, Maat L. Web-based database for facial expression analysis. In: Proceedings of the IEEE international conference on multimedia and expo (ICME); 2005. p. 5.
[5] Lucey P, Cohn J, Kanade T, Saragih J, Ambadar Z, Matthews I. The extended cohn-kanade dataset (ck+): a complete dataset for action unit and emotion-specified expression. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition workshops (CVPRW); 2010. p. 94–101.
[6] Zhao G, Huang X, Taini M, Li S, PietikäInen M. Facial expression recognition from near-infrared videos. Image Vis Comput 2011;29(9):607–19.
[7] Li S, Deng W, Du J. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR); 2017. p. 2852–61.
[8] Goodfellow I, Erhan D, Carrier P, Courville A, Mirza M, Hamner B, et al. Challenges in representation learning: a report on three machine learning contests. In: Neural information processing: 20th international conference, ICONIP 2013, Daegu, Korea, November 3-7, 2013. Proceedings, part III 20; 2013. p. 117–24.
[9] Mollahosseini A, Hasani B, Mahoor M. Affectnet: a database for facial expression, valence, and arousal computing in the wild. IEEE Trans Affect Comput 2017;10(1):18–31.
[10] Aamir M, Ali T, Shaf A, Irfan M, Saleem M. Ml-dcnnet: multi-level deep convolutional neural network for facial expression recognition and intensity estimation. Arab J Sci Eng 2020;45(12):10605–20.
[11] Yan Y, Zhang Z, Chen S, Wang H. Low-resolution facial expression recognition: a filter learning perspective. Signal Process 2020;169:107370.
[12] Sepas-Moghaddam A, Etemad A, Pereira F, Correia PL. Capsfield: light field-based face and expression recognition in the wild using capsule routing. IEEE Trans Image Process 2021;30:2627–42.
[13] Arnaud E, Dapogny A, Bailly K. Thin: throwable information networks and application for facial expression recognition in the wild. IEEE Trans Affect Comput 2022;14(3):2336–48.
[14] Wang K, Peng X, Yang J, Meng D, Qiao Y. Region attention networks for pose and occlusion robust facial expression recognition. IEEE Trans Image Process 2020;29:4057–69.
[15] Cai J, Meng Z, Khan AS, Li Z, O'Reilly J, Tong Y. Probabilistic attribute tree structured convolutional neural networks for facial expression recognition in the wild. IEEE Trans Affect Comput 2022;14(3):1927–41.
[16] Valanarasu JMJ, Oza P, Hacihaliloglu I, Patel VM. Medical transformer: gated axial-attention for medical image segmentation. In: Medical image computing and computer assisted intervention–MICCAI 2021: 24th international conference, Strasbourg, France, September 27–October 1, 2021, proceedings, part I 24. Springer; 2021. p. 36–46.
[17] Liu Z, Wen C, Su Z, Liu S, Sun J, Kong W, et al. Emotion-semantic-aware dual contrastive learning for epistemic emotion identification of learner-generated reviews in moocs. IEEE Trans Neural Netw Learn Syst 2023.
[18] Liu Y, Li G, Lin L. Cross-modal causal relational reasoning for event-level visual question answering. IEEE Trans Pattern Anal Mach Intell 2023;45(10):11624–41.
[19] Cai J, Meng Z, Khan AS, Li Z, O'Reilly J, Tong Y. Island loss for learning discriminative features in facial expression recognition. In: 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018). IEEE; 2018. p. 302–9.
[20] Wen Y, Zhang K, Li Z, Qiao Y. A discriminative feature learning approach for deep face recognition. In: Computer vision–ECCV 2016: 14th European conference, Amsterdam, the Netherlands, October 11–14, 2016, proceedings, part VII 14. Springer; 2016. p. 499–515.
[21] Zhang Y, Wang C, Deng W. Relative uncertainty learning for facial expression recognition. Adv Neural Inf Process Syst 2021;34:17616–27.
[22] Liu P, Lin Y, Meng Z, Lu L, Deng W, Zhou JT, et al. Point adversarial self-mining: a simple method for facial expression recognition. IEEE Trans Cybern 2021;52(12):12649–60.
[23] Jiang J, Deng W. Disentangling identity and pose for facial expression recognition. IEEE Trans Affect Comput 2022;13(4):1868–78.
[24] Liu H, Cai H, Lin Q, Li X, Xiao H. Adaptive multilayer perceptual attention network for facial expression recognition. IEEE Trans Circuits Syst Video Technol 2022;32(9):6253–66.
[25] Wang C, Xue J, Lu K, Yan Y. Light attention embedding for facial expression recognition. IEEE Trans Circuits Syst Video Technol 2021;32(4):1834–47.
[26] Sidhom O, Ghazouani H, Barhoumi W. Three-phases hybrid feature selection for facial expression recognition. J Supercomput 2024;80(6):8094–128.
[27] Sun M, Yan C. Fgenet: a lightweight facial expression recognition algorithm based on fasternet. Signal Image Video Process 2024:1–18.
[28] Feng B, Zhang H. Expression recognition based on visual transformers with novel attentional fusion. J Phys Conf Ser 2024;2868:012036. IOP Publishing.
[29] Zhang F, Chen G, Wang H, Zhang C. Cf-dan: facial-expression recognition based on cross-fusion dual-attention network. Comput Vis Media 2024:1–16.
[30] Tian Y, Zhu J, Yao H, Chen D. Facial expression recognition based on vision transformer with hybrid local attention. Appl Sci 2024;14(15):6471.
[31] Dar T, Javed A, Bourouis S, Hussein HS, Alshazly H. Efficient-swishnet based system for facial emotion recognition. IEEE Access 2022;10:71311–28.
[32] Mohan K, Seal A, Krejcar O, Yazidi A. Fer-net: facial expression recognition using deep neural net. Neural Comput Appl 2021;33(15):9125–36.
[33] Karnati M, Seal A, Yazidi A, Krejcar O. Flepnet: feature level ensemble parallel network for facial expression recognition. IEEE Trans Affect Comput 2022;13(4):2058–70.
[34] Li Q, Han Z, Wu X-M. Deeper insights into graph convolutional networks for semi-supervised learning. In: Proceedings of the AAAI conference on artificial intelligence, vol. 32; 2018.
[35] Wang X, Girshick R, Gupta A, He K. Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018. p. 7794–803.
[36] Woo S, Park J, Lee J-Y, Kweon IS. Cbam: convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV); 2018. p. 3–19.

[37] Xu K. Show, attend and tell: neural image caption generation with visual attention. preprint. arXiv:1502.03044, 2015.
[38] Nair V, Hinton GE. Rectified linear units improve restricted Boltzmann machines. In: Proceedings of the 27th international conference on machine learning (ICML-10); 2010. p. 807–14.
[39] Ioffe S. Batch normalization: accelerating deep network training by reducing internal covariate shift. preprint. arXiv:1502.03167, 2015.
[40] Dosovitskiy A. An image is worth 16x16 words: transformers for image recognition at scale. preprint. arXiv:2010.11929, 2020.
[41] Hendrycks D, Gimpel K. Gaussian error linear units (gelus). preprint. arXiv:1606.08415, 2016.
[42] Xu R, Huang A, Hu Y, Feng X. Gfft: global-local feature fusion transformers for facial expression recognition in the wild. Image Vis Comput 2023;139:104824.
[43] Gong W, La Z, Qian Y, Zhou W. Hybrid attention-aware learning network for facial expression recognition in the wild. Arab J Sci Eng 2024:1–15.
[44] Ma F, Sun B, Li S. Facial expression recognition with visual transformers and attentional selective fusion. IEEE Trans Affect Comput 2021;14(2):1236–48.
[45] Li C, Li X, Wang X, Huang D, Liu Z, Liao L. Fg-agr: fine-grained associative graph representation for facial expression recognition in the wild. IEEE Trans Circuits Syst Video Technol 2023;34(2):882–96.
[46] Chen D, Wen G, Li H, Chen R, Li C. Multi-relations aware network for in-the-wild facial expression recognition. IEEE Trans Circuits Syst Video Technol 2023;33(8):3848–59.

SAMIA NAWAZ YOUSAFZAI received the B.S. degree in software engineering from HITEC University, Taxila, Pakistan, in 2024. She is currently working on multiple projects related to medical imaging, agricultural imaging, sentiment analysis, and fake news classification. Her research interests encompass deep learning, digital image processing, computer vision, and natural language processing.

INZAMAM MASHOOD NASIR received bachelor's, master's, and Ph.D. degrees in computer science from COMSATS University Islamabad, Pakistan, in 2012, 2016, and 2023, respectively. He is currently working on privacy-preserving techniques by embedding blockchain and IoTs with machine-learning techniques for real-world applications. His most recent projects are based on federated learning, explainable AI, and different enhancement techniques to improve the efficiency of models in real-world applications. He is a big fan of nature-inspired algorithms, thus one of his key interests is bio-inspired optimization algorithms. His research interests include machine learning (ML) for medical imaging, agricultural imaging, and hyperspectral imaging.

OUMAIMA SAIDANI received the M.Sc. degree in computer sciences from Paris Dauphine University, France, and the Ph.D. degree in computer sciences from Paris 1-Panthéon Sorbonne University, France. She is currently an Associate Professor with the Information Systems Department, College of Computer and Information Sciences (CCIS-IS), Princess Nourah bint Abdulrahman University (PNU), Saudi Arabia. Her research interests include information systems engineering, business process engineering, IoT, context-aware computing, deep learning, and artificial intelligence.

REFKA GHODHBANI is currently serving as Assistant Professor in the Department of Computer Sciences, Faculty of Computing and Information Technology, Northern Border University, Rafha, Kingdom of Saudi Arabia. She received her Ph.D. degree in computer science and engineering from the Faculty of Sciences, Monastir University, Tunisia, in 2021. Her research interests include real-time image and video processing, embedded system, high level synthesis, machine learning, deep learning, image compression, JPEG2000, FPGA acceleration of image and processing system.

YEONGHYEON GU received his B.S., M.E., and Ph.D. degrees in Computer Science and Engineering from Sejong University, Seoul, South Korea, in 2004, 2006, and 2014, respectively. He is currently the Director of the AI Convergence Research Center and a Professor in the Department of Artificial Intelligence and Data Science at Sejong University, Seoul, South Korea. His research interests include AI, NLP, meta-learning, transfer learning, and related fields

MUHAMMAD SYAFRUDIN received the B.S. degree from UIN Sunan Kalijaga, Yogyakarta, Indonesia, in 2013, and the Ph.D. degree from Dongguk University, Seoul, South Korea, in 2019. He was an Assistant Professor at Dongguk University from 2019 to 2022 and has been an Assistant Professor at Sejong University since 2022. With over ten years of experience in industrial artificial intelligence, data intelligence, industrial analytics, and informatics, his research has been published in journals, such as ESWA, AEJ, COMPAG, Food Control, Sustainable Development, Mathematics, etc. He is listed among Stanford/Elsevier's Top 2% Scientists for 2024 in Artificial Intelligence and Image Processing and serves as a guest editor and a reviewer for Scopus- and SCI-indexed journals.

NORMA LATIF FITRIYANI received the B.S. degree from UIN Sunan Kalijaga, Yogyakarta, Indonesia, in 2014, the M.S. degree from the National Taiwan University of Science and Technology, Taipei, Taiwan, in 2016, and the Ph.D. degree from Dongguk University, Seoul, South Korea, in 2021. She has been an Assistant Professor at Sejong University since 2022. With over seven years of experience in data science, statistical and machine learning, informatics, and image processing, her research has been published in journals such as COMPAG, ESWA, AEJ, Mathematics, Sustainable Development, Food Control, etc. She also serves as a guest editor and reviewer for several international peer-reviewed journals.