IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 4 May 2024, accepted 28 May 2024, date of publication 12 June 2024, date of current version 26 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3413136

RESEARCH ARTICLE

STT-Net: Simplified Temporal Transformer for Emotion Recognition

MUSTAQEEM KHAN[1], (Member, IEEE), ABDULMOTALEB EL SADDIK[1,2], (Fellow, IEEE), MOHAMED DERICHE[3], (Senior Member, IEEE), AND WAIL GUEAIEB[2], (Senior Member, IEEE)

[1]Mohamed Bin Zayed University of Artificial Intelligence (MBZUAI), Abu Dhabi, United Arab Emirates
[2]School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, ON K1N 6N5, Canada
[3]Artificial Intelligence Research Centre (AIRC), Ajman University, Ajman, United Arab Emirates

Corresponding author: Mohamed Deriche (m.deriche@ajman.ac.ae)

This work was supported in part by the Deanship of Research and Graduate Studies at Ajman University under Project 2023-IRG-ENIT 36 and Project 2023-IRG-ENIT 40.

ABSTRACT Emotion recognition is one of the crucial topics in computer vision to efficiently recognize the expression of humans through faces. Recently, transformers have been recognized as a robust architecture, and many vision-based transformer models for emotion recognition have been proposed. The major drawback of such models is the high computational cost of the attention mechanism for computing space-time attention. To that end, we studied temporal feature shifting for frame-wise deep learning models to avoid this burden. In this work, we propose a novel temporal shifting approach for a frame-wise transformer-based model by replacing multi-head self-attention (MSA) with multi-head self/cross-attention (MSCA) to model the temporal interactions between tokens without additional cost. The contextual connection between and inside channels and across time is encoded by the proposed MSCA to enhance the recognition rate and reduce the latency for real-world applications. We extensively evaluated our system on CK+ (Cohn-Kanad) and Fer-2013plus (Facial-Emotion-Recognition) benchmark datasets with geometric transforms-based augmentation to address the imbalance issue in the data. According to the results, the proposed MSCA has either outperformed or closely matched the performance of state-of-the-art (SOTA) techniques. However, we conducted an ablation study on a challenging Fer2013+ dataset to demonstrate the significance and potential of our model for complex emotion recognition tasks.

INDEX TERMS Attention mechanism, deep learning, end-to-end architecture, multi-head self/cross-attention, emotion recognition.

I. INTRODUCTION

Emotion recognition and facial expressions are crucial aspects of non-verbal human communication that directly represent behavior and intentions. However, automatic expression recognition in computer vision is a challenging task due to variations in lighting, environment, and poses [1], [2]. Even humans face difficulties in distinguishing expressions under different conditions [3]. Although recent deep learning-based approaches have significantly improved expression recognition performance, these techniques are still limited to datasets and have limitations in cross-corpus experimentation [4], [5]. The task becomes even more intricate when considering variations induced by lighting, environmental conditions, and posture, making it challenging even for human observers [6], [7].

The Vision Transformer (*ViT*) has been a significant breakthrough in image classification, surpassing traditional deep learning systems [8], [9]. The transformer architecture initially designed for text-based tasks is considered a foundation of *ViT* [10], [11]. Transformers represent an image as patches, following the approach of text transformers, preserving image and token quality while reducing computation. However, optimizing the performance of the *ViT* model

The associate editor coordinating the review of this manuscript and approving it for publication was Jolanta Mizera-Pietraszko.

for emotion recognition requires a nuanced exploration of hyper-parameters [12]. Researchers investigated hyper-parameter tuning to maximize *ViT* model accuracy while probing into the synergy between computational intelligence techniques and their applications [13]. As a result, neuronal architectures that have recently experienced tremendous success in recognition are considered the future of hybrid intelligence [14]. The attention mechanism and their tremendous results in some applications have also made them a potential [15].

The *ViT* model has been widely used in emotion recognition, and researchers have achieved significant success in many tasks in other domains. Still, in computer vision, the transformers have several challenges [16]. Transformers use self-attention mechanisms that are computationally expensive, which makes them challenging to scale to high-resolution visual data. They still consider images as sets of patches rather than grids, losing the inherent spatial relationships that CNNs effectively model. This results in decreased performance on tasks that require strong spatial understanding. Additionally, standard Transformer self-attention has a limited local receptive field. Over-reliance on large pre-trained datasets limits their applicability and causes their performance on fine-grained texture-based problems due to reduced resolution from pacification [17].

This research addresses these limitations by proposing a multi-head self/cross-attention (MSCA) mechanism for human emotion recognition. We introduce an innovative attention mechanism called MSCA as a novel approach to address the limitations of existing front-end reinforcement techniques. Our proposed technique effectively captures temporal interactions between tokens without incurring additional costs. Furthermore, our model uses an attention strategy to establish meaningful relations across tokens, enabling accurate acoustic-to-feature mapping for emotions. Additionally, our model directly incorporates spatial and temporal representations of the input tensor for real-time analysis. The experimental results demonstrate the superior performance of the proposed model on CK+ and Fer2013+ datasets compared to the baseline (See experimental results sections for details). The main contributions are summarized as follows:

- We proposed a mechanism for emotion recognition using multi-headed self/cross-attention (MSCA), which incorporates a temporal shift module to eliminate the need for computing spatial-temporal attention in the multi-channel encoder-decoder attention module.
- We incorporated the concept of cross-attention in MSCA to improve temporal interactions within the transformer block. This is different from token shift, which involves the use of extra shifting modules. As a result, we are able to achieve temporal interaction without adding any computational complexity or requiring any changes to the underlying model architecture.
- Experimental evaluations conducted on the CK+ and Fer2013+ datasets demonstrate that our proposed approach outperforms 3 to 5 percent of the baseline models and the token-shift method in accuracy and 20 percent faster in frame-per-second (fps) processing.

The rest of the article is structured as follows: The background of the domain with recent literature is discussed in Section II, and the proposed methodology is discussed in Section III. The experimental results is described in Section IV, and discussion with comparative analysis are discussed in Section V. Finally, the idea is concluded in Section VI with possible future directions.

II. LITERATURE REVIEW

In effective interpersonal communication, facial expression and emotion play an important role in conveying emotions non-verbally that enhance mutual understanding. In recent years, there have been notable improvements in this area as the importance of precise emotion recognition has gained wider recognition across various application domains [18]. Advancements in computational models and techniques aim to develop technologies with capabilities that begin to approach human-level interpretation of non-verbal signals. This progress shows ongoing efforts to create tools that can understand emotional cues from visual stimuli, just like people do non-verbally in their daily lives [17], [19].

Facial expression/emotion recognition (FER) techniques provide useful insights into human behavioral analysis [20]. Prior FER research primarily focused on manually extracting features from landmarks, textures, and other geometric details [16]. However, recent progress in machine learning and large datasets has advanced computational FER models. These models integrate feature extraction and classification in an end-to-end manner, automating the process and leveraging deep networks' representation [21]. Early approaches applied multi-layer perceptron, support vector machines (SVM), and k-nearest neighbors (KNN) for classification using hand-crafted features like histograms of gradients and eigenvectors. More recent work has proposed robust deep learning-based FER systems [20] that applied principal component analysis (PCA) to reconstruct occluded expressions before extracting Gabor wavelet and geometric features [22]. PCA and linear discriminant analysis (LDA) utilized in these techniques to reduce dimensions before classification. Another study demonstrates noise-resistant recognition through an active contour model for face detection [23], [24]. These models combines two different distance functions to better discriminate faces under variable lighting and identities.

Deep learning frameworks advanced the FER systems instead of conventional methods, which can handle large amounts of data [25], [26]. Our focus is on two recent strategies of deep learning for recognition: Convolution Neural Network (CNN) and Visual Transformer (ViT) - based emotion recognition becouse traditional approaches rely on manual feature extraction, limiting robustness and generalization [27]. Hence, CNNs have gained prominence in the deep learning field for addressing these issues that can learn representations directly from raw image data in an

end-to-end manner without laborious feature engineering [28], [29]. These architectures mimic the human visual cortex through hierarchies of locally connected layers and pooling operations that apply trainable filters to detect patterns across input feature maps [17]. Subsampling layers is used in these networks to reduce the dimensionality while preserving key structure and fully-connected layers enable classification based on global representations [30].

A range of CNN-based FER techniques have achieved state-of-the-art (SoTA) performance [31] by exploring deep 3D and attention-based models to address challenges like micro-expressions and occlusions. Hybrid CNN-SVM approaches leverage extracted CNN embedding for classification [23] and multi-stream, CNN fusion methods integrate temporal and geometric cues [32] to recognized emotions. While CNN excel at automated feature learning, large training datasets and compute requirements remain obstacles. So, the transformers frameworks address these through self-attention, allowing global context modeling more efficiently than CNN, which is used local filters. However, transformers' sequence-based design differs from CNN grid-structured processing of images [33]. Some work advanced deep learning architectures balancing CNN and transformer strengths for robust, scalable facial expression analysis by end-to-end automated feature engineering and efficient modeling of long-range dependencies that help the realization of human-level perceptual capabilities [34].

The Vision Transformer (ViT) was introduced in 2021 with promising results across computer vision tasks like image classification [10]. Inspired by transformers' success in natural language processing, ViT represent images as sequences of patches input to an encoder, learning global representations for classification [11]. Several FER approaches leverage ViT and explore local attention features and global context through squeeze-and-excitation blocks [35]. Similarly, TransFER introduced multi-attention dropping to learn detailed local representations in an adaptive manner, combining ViT and multi-head self-attention [36] for emotion recognition via face images.

Recent work introduces Visual Transformers with Feature Fusion (VTFF) to enhance visual word representations [37] for facial emotion detection. The authors adaptively fuse CNN and local bounding pixel features through attentional feature fusion in VTFF to models contextual information and focuses on discriminative characteristics. Furthermore, another model focuses on low-level grid attention to regularize convolutional filters, and high-level visual transformer attention learns global representation from semantic tokens [38] to efficiently recognize facial emotions. Additionally, a new dataset called Aggregation for ViT on Facial Emotion Recognition (AVFER) was developed that combines training and evaluating ViT configurations for facial expression, which is publicly available [39]. Finally, a novel Squeeze ViT representation technique considers both localized landmark features and global context that address ViT challenges involving parameters and complexity [40].

In the domain of image recognition and classification, as indicated in seminal ViT literature, consistently outperform CNNs in terms of accuracy [10]. However, to enable widespread deployment, ViT architectures must prioritize computational efficiency and scalability without compromising performance. ViT is well-suited to modeling global dependencies across an image via self-attention, excelling at holistic image categorization. However, further enhancing ViT for facial expression recognition necessitates attending to localized changes, especially around the maps, which are most expressive. Therefore, this research proposed developing a novel, optimized architecture called MSCA for emotion recognition to advance the SoTA methods. Rather than solely maximizing classification accuracy, the proposed model will balance global and local feature extraction through strategic architectural design choices and self/cross-attention mechanisms to reduce computational complexity. This aims to leverage MSCA strengths while ameliorating its limitations for the more nuanced task of facial behavior analysis. Parameter optimization will furnish the model with capabilities that generalize robustly across datasets and deployment contexts and are applicable in edge devices for real-time processing.

III. PROPOSED ARCHITECTURE

To maximize the conditional probability $p(y|x)$, our objective is to acquire a mapping across the input sequence channels $\hat{C}$. The input sequence, denoted as $x = (x_1, \ldots, x_j, \ldots, x_c)$, is composed of patches and features. In this context, $x_j \in \mathbb{R}^{(S \times f)}$ represents the feature map of the j^{th} term, encompassing S patches and f features. Conversely, the target sequence $y = (y_1, \ldots, y_k, \ldots, y_u)$ forms a sequence of entire patches with a length of U, where each $y_j \in \mathbb{R}^{(L \times 1)}$ corresponds to the features for the j^{th} term.

The overview and significance of our proposed encoder with baseline transformers architecture is illustrated in **Fig. 1**, where (a) illustrates the original transformer encoder block, incorporating the multi-headed self-attention (MSA) module, and (b) represents the encoder block with token shift, introducing two additional modules to shift intermediate features within the original transformer block. This module executes a single-step forward and backward temporal shift, akin to the token shift strategy. Similarly, (c) showcases the proposed (MSCA) encoder instead of (MSA). Notably, this configuration does not include any additional modules, and the performance is better, with the baseline having reduced latency and cost computations.

A. TOKEN-SHIFTING & TRANSFORMERS

The role of token-shift is briefly investigated in this part based on transformers [41].

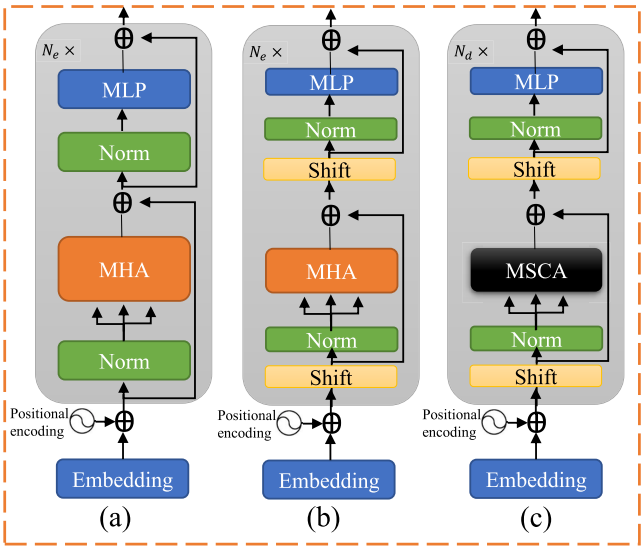

FIGURE 1. (a) Original transformer encoder block, (b) encoder block with token shift, (c) proposed encoder with MSCA block.

1) EMBEDDING OF PATCHES

We consider the input tensor, denoted as $W \in \mathbb{R}^{Se \times H \times W}$. Here, Se represents the number of patches in the images, while H and W represent the dimensions. Each input is divided into the size of patches $ps \times ps$ pixels. These patches are transformed into a tensor $\hat{H} = [\hat{H}_0^1, \ldots, \hat{H}_0^N] \in \mathbb{R}^{W \times N \times D}$, where $\hat{H}_0^j \in \mathbb{R}^{W \times D}$ denotes the i^{th} patch. The total number of patches, denoted by N, is determined as $N = \frac{H \cdot W}{ps^2}$, and each patch has a $D = 3ps^2$ dimension. Subsequently, the input patch $\hat{H}0^i$ undergoes a transformation using an embedding matrix $En \in \mathbb{R}^D$ and positional encoding En_{pos}. The transformation process is as follows:

$$x_0 = (c_0, x_0^1 \, En, x_0^2 \, En, \ldots, x_0^N En) + En_{pos} \tag{1}$$

In Eq. 1, $c_0 \in \mathbb{R}^{W \times D}$ represents a token of the class. The embedding of patch $B_0 \in \mathbb{R}^{W \times (n+1) \times D}$ is obtained by adding the product of each x_0^i and the embedding matrix En (Encoder) to the class token c_0. The positional encoding En_{pos} is added to x_0. The patch embedding B_0 is initially utilized as an input for the encoder block.

2) ENCODER BLOCK OF BASELINE

The encoder block is a crucial component in the Transformer architecture (baseline), designed for parallelized processing of input sequences by incorporating the MSA mechanism, where the input is divided into heads, allowing the model to capture different relationships within the sequence. The scaled dot-product attention is used to calculate the attention scores and then passes through a series of operations to process the information at each position independently further. We modify this hierarchical structure, consisting of self/cross attention MSCA and feed-forward layers with residual connections, which enable the encoders to capture intricate patterns and dependencies in input, and multiple MSCA blocks are typically stacked to enhance the model's representative capacity. The visual flow diagram of the baseline encoder is illustrated in **Fig. 1(a)**, and working mechanism is as follow: Let B_l denote the input sequence up to l^{th} term in the encoder. The resulting output B_l from the blocks can be represented as

$$B'_l = \text{MSA}\,(Norm(B_{l-1})) + B_{l-1} \tag{2}$$

$$B_l = \text{MLP}\,\left(Norm(B'_l)\right) + B_l^1 \tag{3}$$

In Eq. 2 and Eq. 3, the input B_l is passed through the MSA module, followed by the normalization layer and an element-wise addition with B_{l-1}. The resulting output is denoted as B'_l. Subsequently, B'_l undergoes Linear normalization denoted by *Norm* and is fed into an MLP. The output of the MLP is added element-wise with B_l^1, resulting in the final output B_l of the l^{th} encoder block.

3) SHIFT-MODULES

Vision Transformers have achieved remarkable success with self-attention but struggle to capture fine-grained local context due to the loss of 2D positional information during patchification. The token-shift mechanism aims to address this limitation by incorporating a notion of locality into the self-attention computation. It augments token embedding with 2D positional encoding before self-attention. Rather than attending to tokens directly, the query is shifted to nearby patches within a neighborhood window. This effectively makes the attention map more focused on local regions, enabling *ViTs* to better model fine-grained relationships between neighboring patches in a manner analogous to convolutional kernels. Token-shift enables Transformers to balance their strengths in global context modeling through self-attention with stronger localization abilities, akin to CNNs, and achieve improved performance for dense prediction tasks while maintaining competitive data efficiency.

The visual flow diagram of the baseline encoder with shift module is illustrated in **Fig. 1(b)** with working strategy. Therefore, the proposed MSCA attention block includes two shift modules representing the token-shift mechanism in **Fig. 1(c)**. The shift operations can be described as

$$B'_{(l-1)} = \text{shift}(B_{(l-1)}) \tag{4}$$

$$B''_l = \text{MSA}\,\left(Norm(B'_{l-1})\right) + B'_{l-1} \tag{5}$$

$$B'''_l = \text{shift}(B''_l) \tag{6}$$

$$B_l = \text{MLP}\,\left(Norm(B'''_l)\right) + B'''_l \tag{7}$$

In this process, the incoming tensor $B_{in} \in \mathbb{R}^{(W \times (N+1) \times D)}$ and generate a result B_{out} in similar dimensions. These modules shift the section of B_{in} that token corresponds to class $(B_{(in,T,O,D)})$ to the beginning of the other part of B_{in}

thereby keeping the second sections unaltered.

$$B_{\text{out},T,O,D} = \begin{cases} B_{\text{in},T-1,O,D} \\ B_{\text{in},T+1,O,D} + D_f \\ B_{\text{in},T,O,D} \end{cases} \tag{8}$$

$$B_{\text{out},T,N,D} = B_{\text{in},T,N,D} \tag{9}$$

These shift operations are part of the token-shift mechanism employed in the proposed attention block.

B. PROPOSED MSCA ATTENTION MECHANISM

This section discusses the main differences between the MSA (multi-headed self-attention) and the proposed MSCA (multi-headed self/cross-attention) networks. We provide an overview and the main difference between MSA and proposed MSCA in terms of theoretical background and a step-by-step procedure for building and working strategy with input data. Relations of Eq. 2 and Eq. 3 demonstrate the MSA block of baseline encoders that we propose here to replace by MSCA. A step-by-step procedure and strategy to show the main distinction between the baseline MSA and the proposed MSCA networks is explained below.

The MSA module computes the query ($Q^{(t)}$), value ($V^{(t)}$), and key ($K^{(t)}$) with a certain part ($B^{(t)} \in \mathbb{R}^{(N+1)\times D}$) of the input feature ($B \in \mathbb{R}^{(W\times(N+1)\times D)}$) at a given Number N. This process is carried out using the following expressions:

$$Q^{(t)}, V^{(t)}, K^{(t)} = B^{(N)} \cdot (w_k, w_q, w_v) \tag{10}$$

In the given equations, the matrices $w_k, w_q, w_v \in \mathbb{R}^{(D\times N)}$ represent the embedding matrices, and the input feature B is composed of patches $B^{(1)}, \ldots, B^{(N)}$. These values are utilized to calculate the attention of the i^{th} head:

$$\mathrm{H}_i^{(t)} = P(Q_i^t, K_i^t)V_i^t \in \mathbb{R}^{(N+1)\times \frac{D}{H}} \tag{11}$$

$$P(Q, K) = \text{softmax}\left(\frac{QK^{t'}}{\sqrt{D}}\right) \tag{12}$$

In Eq. 11 and Eq. 12, the patch P is considered, where $Q_i^t \in R^{(N+1)}$ represents a part of the Q^t heads denoted by $Q^{(t)} = (Q_1^t, \ldots, Q_i^t, \ldots, Q_H^P)$, with H denoting the head and t' representing the transpose.

$$\text{MSA}(B^{(N)}) = [\mathrm{H}_1^N, \ldots, \mathrm{H}_H^N] \tag{13}$$

where K_i^t and V_i^t are the same heads stacked from the MSA, where patches in the i^{th} head attend to another patch within a similar pitch. This implies that no temporal interactions are occurring between the different patches. The visual framework of the proposed MSCA architecture is illustrated in **Fig.2** with related components, and the description of each block is discussed in the upcoming sections.

1) KEY & VALUE KV POSITION IN PROPOSED MSCA MODULE

The designed MSCA module incorporates saliency across the input, allowing patches in the i^{th} tensor at level $N-1$ and $N+1$. It's achieved through shift operations, where the query Q, key K, and value V are generated for each tensor (input), and depending on the chosen configuration, K and V can be shifted, allowing the current frame's query to attend pairs of key values in other pitch in the proposed MSCA mechanism, which is described as follows:

$$\text{head}_j^t = \begin{cases} I(Q_j^t, K_j^{t-1})V_j^{t-1}, & 1 \le j < h_b, \\ I(Q_j^t, K_j^{t-1})V_j^{t+1}, & h_b \le j < h_b + h_f, \\ I(Q_j^t, K_j^{t-1})V_j^{t-1}, & h_b + h_f \le j < h. \end{cases} \tag{14}$$

In Eq. 14, the initial heads experience a shift backward, the following heads h_f undergo a shift forward, and the remaining heads do not shift. This approach is referred to as $MSCA-KV$, where I represents the input image/pitch. Initially, keys, queries, and values are calculated for each input, and subsequently, some of these values are shifted before the attention computation step. The solid arrows indicated shifts interval $t+1$ and $t-1$ for key-value. The same shift process occurs simultaneously in all other input tensors, which is illustrated in **Fig.3**.

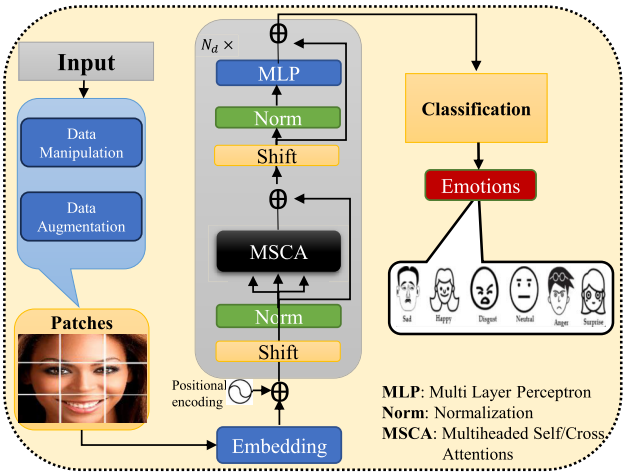

FIGURE 2. A visual illustration of the proposed MSCA model for emotion recognition.

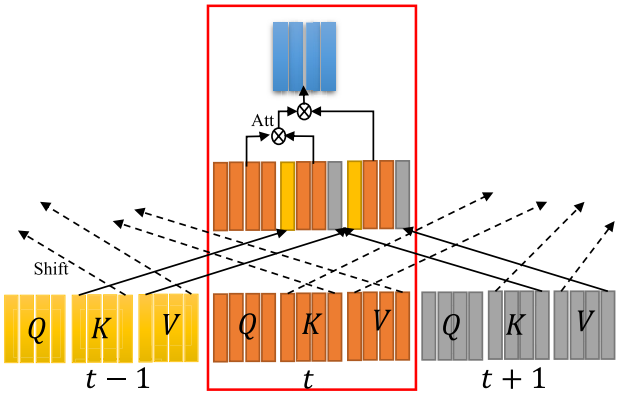

FIGURE 3. A visual illustration of the proposed MSCA-KV learning module and attention mechanism for emotion recognition.

2) VALUE V POSITION IN PROPOSED MSCA MODULE

The other possible solutions in a shift module for the proposed MSCA involve shifting only the value V while keeping the query Q and key K unchanged, which is visually illustrated in **Fig. 4**. The corresponding mathematical concept is formally modeled by Eq. 15.

$$\text{head}_i^t = \begin{cases} I(Q_i^t, K_i^t)V_i^{t-1}, & 1 \le j < h_b \\ I(Q_i^t, K_i^t)V_i^{t+1}, & h_b \le j < h_b + h_f \\ I(Q_i^t, K_i^t)V_i^t, & h_b + h_f \le j \le h \end{cases} \tag{15}$$

This approach, known as MSCA-V, differs from the previous shift method as only the value component is shifted, while the query and key components remain within their respective place. It's important to note that this approach may not be commonly used because it separates the key and value from different pitches. Apart from V in MSCA, there are some additional combinations, and all variants will be examined in the experiments to evaluate their performance. Notably, QKV in MSCA can be regarded as a form of feature shifting, as the attended features are computed within each input and subsequently shifted accordingly.

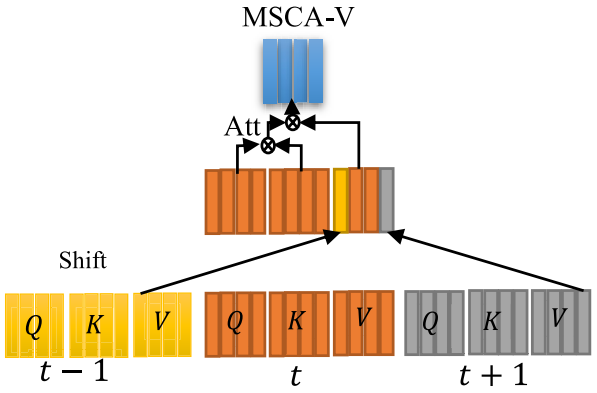

FIGURE 4. A visual illustration of the proposed MSCA-V learning module and attention mechanism for emotion recognition.

3) PITCH KEYS & VALUES (*PKV*) IN PROPOSED MSCA MODULE

All MSCA variants incorporate shift operations in the head dimension D, and it is also possible to apply similar shift variants in the patch dimension $N + 1$. The shapes of K^t, Q^t, and V^t are $\mathbb{R}^{(N+1)\times D}$, where the first dimension represents patches and the second dimension represents heads. This allows for shifts in both the head and patch dimensions. Initially, the keys and values are organized as stacks, each consisting of keys and values from different frames of patches. For example, $K^t = (K_1^t, \ldots, K_1^t, \ldots, K_N^t)$ and $V^t = (V_0^t, \ldots, V_1^t, \ldots, V_N^t)$. The shift operations for the keys are defined as

$$K_N^{'t} = \begin{cases} K_N^{t-1}, & 1 \le N < N_b \\ K_N^{t+1}, & N_b \le N < N_b + N_f \\ K_N^t, & N_b + N_f \le N < N \end{cases} \tag{16}$$

TABLE 1. Statistical analysis of the Fer-2013plus and CK+ datasets.

Classes/Emotions	Label	Fer-2013+ Samples	CK+ Samples
Anger	0	2656	135
Happy	1	9038	207
Sad	2	3752	84
Surprise	3	3941	249
Fear	4	682	75
Disgust	5	157	177
Neutral	7	10996	54

where K_N^t and $V_N^t \in \mathbb{R}^D$ represent the values and keys of patches at interval T. The key of certain parts in the current pitch is transferred to K' and V' using the same approach. Finally, the computation of the i^{th} head is performed following Eq. 17.

$$\text{H}_i^t = I(Q_i^t, K_i^{'t})V_i^{'t} \Rightarrow (Q_i^t, K_i^t)V_i^{'t} \tag{17}$$

We refer to the proposed version of MSCA with patch shift using keys and values as MSCA-PKV and the version with patch shift using only values as MSCA-PV. These variants are similar to MSCA-V (see **Fig. 4**, which involves shifting in the patch direction, focusing on the value component. Similarly, there are seven total variants, including MSCA-PV and so on.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

A. DATASETS & PRE-PROCESSING

In this study, we used two benchmark datasets, Fer-2013plus [42] and Cohn-Kanade (CK+) [43], to evaluate the performance and robustness of the proposed system.

Fer-2013plus [42]: The Fer-2013plus dataset is created in 2016 as an extension of the FER-2013 dataset provided by Kaggle. The images in this dataset are greyscale facial expressions measuring 48 pixels by 48 pixels that have been classified into various emotion categories. The Fer-2013plus dataset contains 35,887 images representing seven facial expressions, with labels ranging from zero to six. The distribution of images across these classes varies from each other as illustrated in **Table 1**. Test sets consist of 7178 samples split between public and private, whereas training sets consist of 28,709 samples. We have validated our approach on this standardized facial expression/emotion recognition benchmark with a variety of emotional categories, and the testing results visually showed in **Fig. 7**.

CK+ [43]: This dataset is the extended version of the original Cohn-Kanade (CK) dataset, which consists of 593 video sequences and labeled images collected in a controlled laboratory environment. The data includes 123 subjects ranging from 18 to 30 years old, and the resolution of the images is 640 × 480, and 640 × 490 pixels at 8-bit grayscale version. In addition to the six basic emotions of anger, disgust, fear, happiness, sadness, and surprise, the CK+ dataset also includes the emotion of neutral as shown in **Table 1**. For model evaluation, we utilized 80% of the data for training and 20% for validation purposes. This dataset provides a standardized resource to benchmark

FER algorithms incorporating a wider range of emotional categories than prior datasets.

Pre-processing: In pre-processing, the data augmentation technique is applied to enhance and enrich the datasets and reduce overfitting for robust model training. All images are undergoing multiple transformations aimed at expanding the available training samples in a computationally efficient manner, and scales are adjusted to normalize the size variation of each image in the dataset. Random rotations are applied to artificially introduce angular changes in orientation and flipped horizontally and vertically to generate left-right and up-down variants from existing images. Additionally, Pixel values are normalized to facilitate model convergence by ensuring input features lie within a similar distribution and scale to help in the training process for more efficient outcomes. The visual representation module of data pre-processing is illustrated in **Fig. 2**.

B. RESULTS AND ANALYSIS

We used the accuracy matrix defined in Eq. 18 to evaluate the effectiveness of our model and conduct a comparative analysis with baseline models.

$$\text{ACC} = \frac{\sum_{i=0}^{6} T_i}{\sum_{i=0}^{6} T_i + F_i} \tag{18}$$

Similarly, F_i represents the number of predictions of samples of the i^{th} class that do not match that class. Dataset classes are associated with i.

This matrix indicates the model performance across all classes and defines the ratio between the proposed model correct predictions and the total number of predictions. Additionally, a K-fold cross-validation technique is used to extensively evaluated the proposed model to report the mature results. During experimentation, a 10-fold cross-validation approach is adopted whereby the data is split into ten folds, nine used for training and one for validation in each iteration. The procedure involves random shuffling, splitting into folds, training testing and saving the evaluation score on each iteration. Furthermore, the model is tested on a separate hold-out test set to obtain an overall accuracy matrix and compared with SoTA in **Table 4**.

C. ABLATION STUDY

Overall, the baseline transformers and our proposed multi-head self/cross-attention (MSCA) have $d_{\text{ff}} = 1024$ hidden neurons, and $h = 3$ heads at initial version. During configuration, we analyse various setups with MSCA having $h = 4$, via other transformers, as stated in **Table 2**, with equivalent model sizes utilizing Fer-2013plus dataset. The accuracy demonstrates the outcomes of all experiments, which can be calculated for different mechanisms. The larger accuracy value shows better performance with similar setup and input data. **Table 2** demonstrates that the "Proposed MSCA" model outperforms the other approaches in terms of accuracy for emotion recognition. The proposed setup achieves better results with the rest using similar dataset. Conclusively, we performed an ablation study to illustrate the significance of various attention layers and train different types of transformers to select the best architecture for emotion recognition.

TABLE 2. Ablation study of the proposed system with different learning strategies using Fer-2013plus dataset using image as input to the model.

Architecture	Accuracy (%)
Convolution Neural Network (CNN)	76.50
CNN + Self-Attention mechanism	80.00
Vision transformer (ViT)	80.32
ViT + Token-shift	80.68
ViT + Self Attention	83.00
ViT + Cross Attention	84.21
ViT + Self/Cross Attention	86.50
ViT + Token-shift + Self/Cross Attention	**93.20**

TABLE 3. Classification scores of MSCA on Fer-2013plus dataset.

Emotion/Class	Precision	Recall	F1-Score
Anger	0.83	0.85	0.84
Disgust	0.82	0.80	0.81
Fear	0.92	0.90	0.91
Happy	0.90	0.94	0.92
Sad	0.83	0.81	0.82
Surprise	0.89	0.90	0.90
Neutral	0.94	0.93	0.93
Accuracy		**0.95**	
Macro Average	0.89	0.95	0.92
Weighted Average	0.94	0.94	0.95

D. QUANTITATIVE ANALYSIS

Fer-2013plus: This dataset includes images representing seven basic facial expressions as depicted in the confusion matrix **Fig. 5**, it appears that most classes have been predicted accurately during validation except for fear expressions. For some samples, the predicted class of fear is incorrectly given as anger and disgust. Still, the overall classification accuracy is better than the baseline as reflected by the precision, recall, and F1-score metrics calculated for each sample in the test set as shown in **Table 3**. Furthermore, **Fig. 5** depicts the matrix for the Fer-2013plus dataset, providing insight into how well the model differentiated between the true versus predicted expression categories during evaluation. The matrix visualization helps analyse what types of errors or misclassifications occurred across expressions. The diagonal values in the confusion matrix is showing the actual prediction corresponding each class, which is highlighted in **Fig. 5**.

CK+: For model significance and robustness, we trained and tested the proposed model utilizing CK+ dataset as well. During the model testing, our system correctly recognized the most expression classes and identified them with a high precision rate. However, similar to findings with the prior dataset, Sad expression was occasionally misclassified as neutral. Still, this misclassification could be attributed to visual similarities between the tension displayed in sad and neutral as depicted in **Fig. 6**. Both involve downward-turned

TABLE 4. Comparison of the proposed MSCA model with existing models on the Fer-2013plus and CK+ datasets.

Method	Year	Architecture	Classifier	Parameters	Fer-2013plus (%)	CK+(%)
Dosovitskiy et al. [10]	2021	ViT model	—	12 million	73.36	96.21
Xue et al. [36]	2021	Trans-FER model	NN	12 million	90.83	98.80
Ma et al. [37]	2021	ViT + Feature Fusion	—	21.3 million	88.81	—
Huang et al. [38]	2021	FerVT	—	2.6 million	90.04	99.00
Kim et al. [44]	2019	CNN	SVM	2.4 million	73.73	—
Georgescu et al. [45]	2019	CNN-NE	SVM	8.0 million	87.76	—
Kim et al. [40]	2022	Squeeze ViT	NN	3.48 million	—	99.54
Aouayeb et al. [35]	2021	ViT + SE	NN	—	—	98.80
Wu et al. [46]	2023	FER + CHC	SVM	5.0 million	90.81	99.49
Meena et al. [47]	2023	DCNN	NN	5.0 million	79.00	95.00
Kumari et al. [48]	2023	ConvNet	—	7.0 million	82.7	97.00
Kumar et al. [49]	2023	CNN + LPQ/LBP	MSVM	11.0 million	—	94.20
Boughanem et al. [50]	2023	MCNN	NN	—	94.02	98.80
Our Method	**2024**	**Proposed MSCA**	**NN**	**2.2 million**	**95.12**	**98.30**

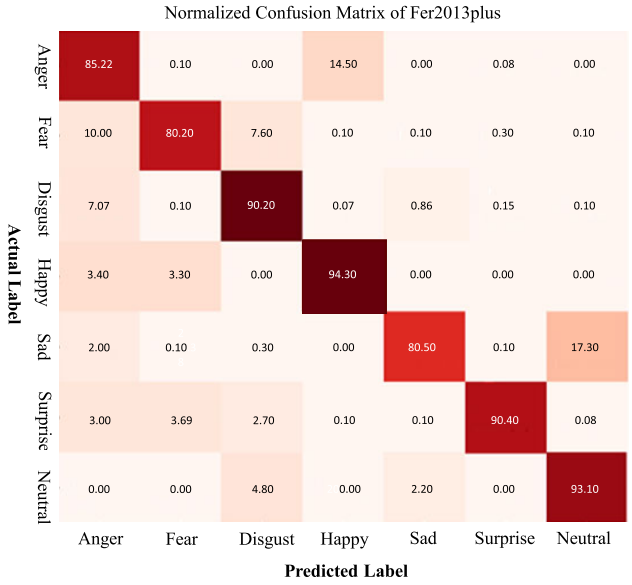

FIGURE 5. Proposed model prediction results among actual and predicted labels over the Fer-2013plus dataset. The diagonal value represents the actual recall across each class.

mouths and eyes, which may sometimes confound automated models and confuse the model to correctly predicted the emotion. The accuracy score incorporates precision, recall, and F1 metrics calculated individually for each class calculated and mentioned in **Table 5**. This provides a more nuanced view of how well different emotions are predicted beyond just an aggregate accuracy percentage. With an overall accuracy of 95%, most expressions were correctly classified most of the time. However, exploring performance on a per-class basis through the confusion matrix sheds light on where the model may require targeted improvements to better discern emotionally similar but distinct facial cues like fear versus sadness. The diagonal values in the confusion matrix is showing the actual prediction corresponding each class, which is highlighted in **Fig. 6**.

E. QUALITATIVE ANALYSIS

We tested our proposed system over different emotions and reported their confidence score, which is shown in **Fig. 7**,

TABLE 5. Classification scores of MSCA on CK+ dataset.

Score of MSCA on CK+			
Emotion/Class	**Precision**	**Recall**	**F1-Score**
Anger	0.99	0.96	0.98
Disgust	1.00	0.96	0.97
Fear	0.98	0.97	0.98
Happy	1.00	0.95	0.96
Sad	0.90	0.86	0.90
Surprise	0.98	0.95	0.97
Neutral	1.00	0.97	0.98
Accuracy		**0.98**	
Macro Average	0.98	0.97	0.97
Weighted Average	0.95	0.98	0.98

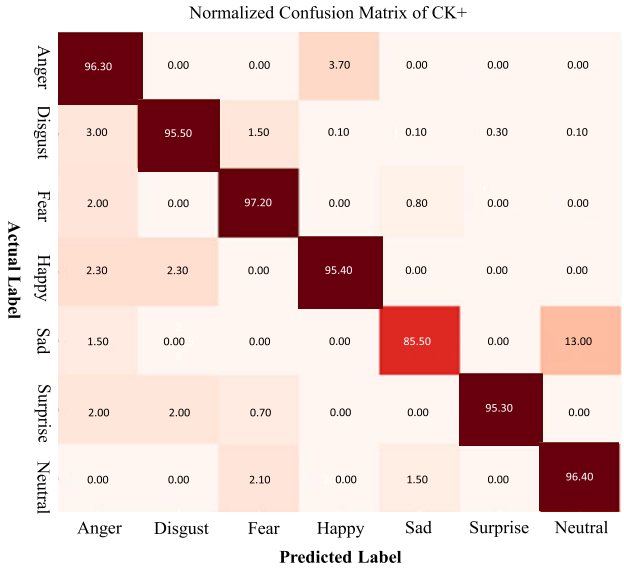

FIGURE 6. Proposed model prediction results among actual and predicted labels over CK+ dataset. The diagonal value represents the actual recall across each class.

where (a) and (b) show the testing performance of the facial emotion recognition model on the Fer-2013plus and CK+ datasets, respectively, by displaying the accuracy across different probability thresholds. In **Fig. 7 (a)**, the model achieves over 95% accuracy in classifying anger, disgust, and

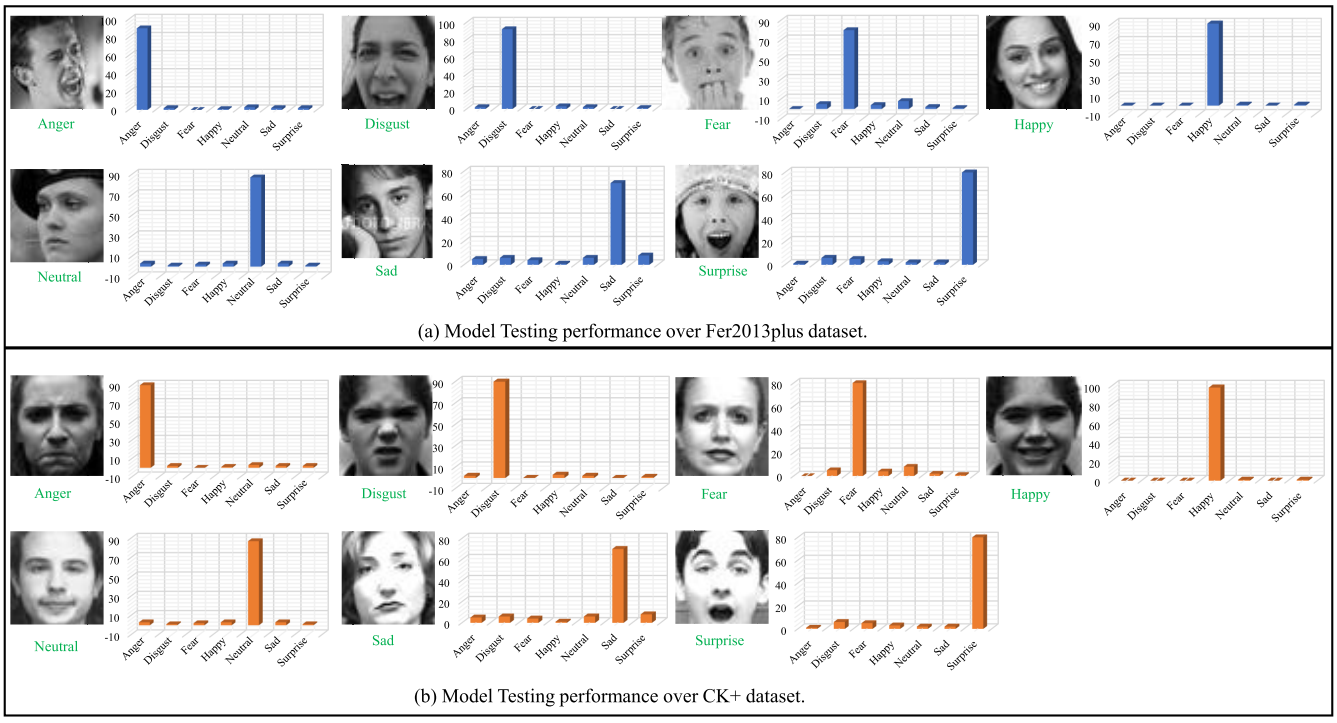

FIGURE 7. Performance of the proposed MSCA model on different emotions along with confidence scores.

sadness across all thresholds when tested on Fer-2013plus data. However, the accuracy for fear decreases significantly at lower thresholds. **Fig. 7 (b)** depicts the performance on the CK+ dataset, where accuracy remains above 95% for classifying anger, disgust, and surprise but drops noticeably for sad, surprise, and fear emotions as the threshold is reduced. Thus, the visual results indicate that while the model robustly classifies certain expressions like anger and disgust, it struggles more with emotionally ambiguous categories such as fear and sadness that involve subtle facial differences. The accuracy is also dependent on the threshold choice, highlighting the need to optimize the parameter for the best performance for a high confidence score. Overall, the proposed model confidence score is better than the baseline quantitative and qualitative analysis. The visual illustration of each emotion and the model score is represented in **Fig. 7**.

V. DISCUSSION AND COMPARATIVE ANALYSIS

The provided **Table 4** compares various methods of FER, showcasing the architecture and accuracy of each model on the Fer-2013plus and CK+ datasets. Each row represents a different method, employing diverse architectures of deep and machine learning algorithms. The accuracy percentages are reported for both datasets, with CK+ generally exhibiting higher accuracy. Notably, the last row shows the proposed method MSCA architecture results, achieving 95.12% accuracy on Fer-2013plus and 98.30% on CK+, indicating competitive performance compared to the other methods in the **Table 4** as well as the computation cost of the designed model is quite reasonable for edge devices in real-time, which is mentioned in **Table 6** and visually illustrated in **Fig. 8** utilizing different architectures. The results highlight the potential effectiveness of the proposed MSCA model in emotion recognition tasks with reduced latency time with higher frame per second (FPS) rate.

Our proposed MSCA model achieved better results against recent work that utilized various deep learning architectures ranging from basic CNNs to memory networks, fine-tuning and generative models, and vision transformers as mentioned in **Table 4**. An important finding is that not all studies report results on both datasets, limiting direct cross-method assessment. Nonetheless, it can be seen that deeper CNN models like *ViT* and fine-tuning approaches have realized the highest accuracy with computationally expensive models. Significantly, the proposed MSCA model outperforms all prior works in terms of accuracy and computation as mentioned in **Table 4 & 6**. This comparison aims to assess the performance of the MSCA model in relation to established works, providing insights into its effectiveness and potential advancements in emotion recognition tasks across diverse datasets. Overall, the tables provide a useful quantitative summary to position the new model within the SoTA, though human-level recognition ability still remains elusively above current techniques.

A. COMPUTATIONAL ANALYSIS FOR EDGE DEVICES

We optimize the proposed MSCA model to resource-constrained edge devices, it is crucial to optimize the models

TABLE 6. Performance comparison of the proposed MSCA model using different frameworks on the Jetson Xavier board.

Framework	Format	Frame per Second (FPS)			Model Size (MB)
		CPU	GPU	Jetson	
Keras	FP32	4.50	7.00	3.00	88.00
TF-Lite	FP32	4.85	8.90	2.26	88.00
PyTorch	FP32	6.00	9.00	5.50	83.80
ONNX	FP32	11.45	13.00	10.30	44.50
ONNX	FP16	11.90	14.50	11.00	45.30
ONNX	FP08	13.00	14.90	11.50	40.00
Tensor-RT	FP32	17.00	21.90	30.00	50.40
Tensor-RT	FP16	17.50	20.00	36.50	50.40
Tensor-RT	**FP08**	**28.00**	**25.00**	**48.50**	**34.00**

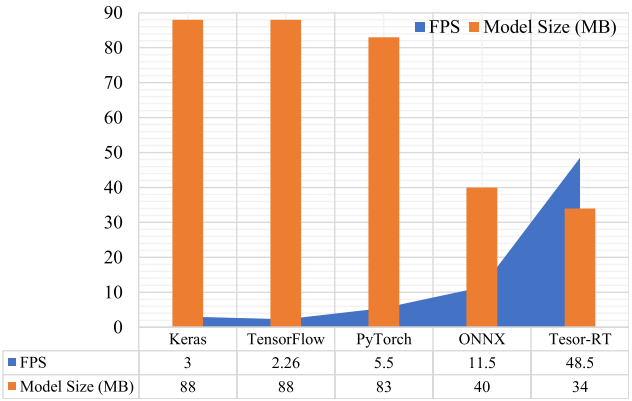

FIGURE 8. Effect of different frameworks (Keras, TensorFlow, PyTorch, ONNX, and Tensor-RT) on the model size and frames per second (FPS).

for low latency, small memory footprint and high energy-efficiency. We employed several techniques during the model optimization process. First, model compression methods like pruning, quantization and distillation are used to shrink proposed models into a more compressed form without loss in accuracy. Its drastically reduces the memory and storage requirements for the proposed model to deploy in real-time applications. Further, model parameters and weights are encoded efficiently using standards ONNX (open neural network exchange) technique. Operators within the model graph are fused together to minimize computational operations. Additionally, we optimize the model to lower precision numeric formats like FP08 (floating point) to more accelerate the model speed without loss any accuracy. The end goal is to produce a model that delivers fast and energy-efficient inference while preserving good prediction quality for deployment on bandwidth and resource-constrained edge devices. The detail experimental results of optimized model is shown in **Table 6** and visually illustrated in **Fig. 8** to show the high FPS over various frameworks.

In **Fig 8**, we shows the FPS rate and model size of the proposed system over different frameworks such as Keras, TensorFlow, PyTorch, open neural network exchange (ONNX) and Tensor-RT. Keras and TensorFlow have the lowest FPS rate of 3 and 2.26 respectively, despite having the largest model sizes of 88MB and PyTorch achieves a bit higher FPS rate of 5.5 with a slightly smaller model size of 83 MB, showing it has better optimizations than Keras/TF. Furthermore, we optimize our model by ONNX, which provides a significant boost in FPS rate to 11.5 while further reducing the model size that shows the runtime performs. Hence, we convert our model to Tensor-RT that achieves the highest FPS rate of 48.5 with the smallest model size of because its a dedicated inference optimization framework so it is able to optimize the model through techniques like operator fusion, kernel auto-tuning, tensor cores etc to maximize throughput. In conclusion, frameworks with dedicated focus on runtime optimizations like ONNX and TensorRT are able to achieve much higher efficiencies in terms of both speed and size for the same model compared to general ML frameworks. Finally, our optimized Tensor-RT model is ready to deploy on edge-devices for real-time applications.

VI. CONCLUSION

The proposed research introduces a novel approach to facial expression recognition (FER) using a deep learning-based method that leverages the multi-head self/cross-attention (MSCA) mechanism within a transformer architecture. This approach aims to improve FER performance across different datasets while being optimized for edge devices. Our experiments demonstrate that the MSCA-based method outperforms 3 to 5 percent baseline models in terms of accuracy and 20 percent in latency and can be easily adapted for real-time applications with minimal changes to model parameters. The experiments were conducted on the Fer-2013plus and CK+ datasets using a consistent custom structure, as well as with variations in the MSCA configuration. Future research could explore unsupervised pre-training techniques and further optimize pre-processing, feature extraction, and dataset balancing to enhance the FER system's efficiency. Despite significant progress, there is still room for improvement in creating sustainable and publicly accessible FER systems. Incorporating more advanced AI and ML techniques could further enhance the FER system's capabilities.

ACKNOWLEDGMENT

The authors acknowledge the invaluable contribution of artificial intelligence (AI) tools in enhancing the efficiency and advancements in their research endeavors.

REFERENCES

[1] Y. R. Veeranki, L. R. M. Diaz, R. Swaminathan, and H. F. Posada-Quintero, "Nonlinear signal processing methods for automatic emotion recognition using electrodermal activity," *IEEE Sensors J.*, vol. 24, no. 6, pp. 8079–8093, Mar. 2024.

[2] P. A. Gavade, V. S. Bhat, and J. Pujari, "Improved deep generative adversarial network with illuminant invariant local binary pattern features for facial expression recognition," *Comput. Methods Biomechanics Biomed. Eng., Imag. Visualizat.*, vol. 11, no. 3, pp. 678–695, May 2023.

[3] L. Zongxing, H. Baizheng, C. Yingjie, C. Bingxing, Y. Ligang, H. Haibin, and L. Zhoujie, "Human–Machine interaction technology for simultaneous gesture recognition and force assessment: A review," *IEEE Sensors J.*, vol. 23, no. 22, pp. 26981–26996, Nov. 2023.

[4] T. Lin, Y. Wang, X. Liu, and X. Qiu, "A survey of transformers," *AI Open*, vol. 3, pp. 111–132, Sep. 2022.
[5] X. Liu, X. Cheng, and K. Lee, "GA-SVM-Based facial emotion recognition using facial geometric features," *IEEE Sensors J.*, vol. 21, no. 10, pp. 11532–11542, May 2021.
[6] Y. Liu, Y. Zhang, Y. Wang, F. Hou, J. Yuan, J. Tian, Y. Zhang, Z. Shi, J. Fan, and Z. He, "A survey of visual transformers," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 6, pp. 1–21, Jun. 2024.
[7] J. Chaki, N. Dey, F. Shi, and R. S. Sherratt, "Pattern mining approaches used in sensor-based biometric recognition: A review," *IEEE Sensors J.*, vol. 19, no. 10, pp. 3569–3580, May 2019.
[8] C. R. Chen, Q. Fan, and R. Panda, "CrossViT: Cross-attention multi-scale vision transformer for image classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 347–356.
[9] R. Krishna, K. Das, H. K. Meena, and R. B. Pachori, "Spectral graph wavelet transform-based feature representation for automated classification of emotions from EEG signal," *IEEE Sensors J.*, vol. 23, no. 24, pp. 31229–31236, Dec. 2023.
[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
[11] L. Tunstall, L. von Werra, and T. Wolf, *Natural Language Processing With Transformers*. Sebastopol, CA, USA: O'Reilly Media, Inc., 2022.
[12] Y. Xu, H. Wei, M. Lin, Y. Deng, K. Sheng, M. Zhang, F. Tang, W. Dong, F. Huang, and C. Xu, "Transformers in computational visual media: A survey," *Comput. Vis. Media*, vol. 8, no. 1, pp. 33–62, Mar. 2022.
[13] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer, "Scaling vision transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1204–1213.
[14] X. Shi, H. Gu, and B. Yao, "Fuzzy Bayesian network fault diagnosis method based on fault tree for coal mine drainage system," *IEEE Sensors J.*, vol. 24, no. 6, pp. 7537–7547, Mar. 2024.
[15] P. D. M. Fernandez, F. A. G. Peña, T. I. Ren, and A. Cunha, "FERAtt: Facial expression recognition with attention net," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 837–846.
[16] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, Z. Yang, Y. Zhang, and D. Tao, "A survey on vision transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 87–110, Jan. 2023.
[17] N. Aloysius and M. Geetha, "A review on deep convolutional neural networks," in *Proc. Int. Conf. Commun. Signal Process. (ICCSP)*, Apr. 2017, pp. 0588–0592.
[18] P. Naga, S. D. Marri, and R. Borreo, "Facial emotion recognition methods, datasets and technologies: A literature survey," *Mater. Today, Proc.*, vol. 80, pp. 2824–2828, Oct. 2023.
[19] S. Li, C. Wang, B. Yang, X. Liang, and J. Li, "An effective recognition method for particle coincidence in double differential impedance cytometry," *IEEE Sensors J.*, vol. 23, no. 16, pp. 18070–18080, Aug. 2023.
[20] M. Sharif, F. Naz, M. Yasmin, M. Shahid, and A. Rehman, "Face recognition: A survey," *J. Eng. Sci. Technol. Rev.*, vol. 10, no. 2, pp. 166–177, 2017.
[21] Z. Song, "Facial expression emotion recognition model integrating philosophy and machine learning theory," *Frontiers Psychol.*, vol. 12, pp. 1–19, Sep. 2021.
[22] J. Mahata and A. Phadikar, "Recent advances in human behaviour understanding: A survey," *Devices Integr. Circuit.*, vol. 1, no. 1, pp. 751–755, Mar. 2017.
[23] D. D. Sawat and R. S. Hegadi, "Unconstrained face detection: A deep learning and machine learning combined approach," *CSI Trans. ICT*, vol. 5, no. 2, pp. 195–199, Jun. 2017.
[24] M. H. Siddiqi, R. Ali, A. M. Khan, E. S. Kim, G. J. Kim, and S. Lee, "Facial expression recognition using active contour-based face detection, facial movement-based feature extraction, and non-linear feature selection," *Multimedia Syst.*, vol. 21, no. 6, pp. 541–555, Nov. 2015.
[25] Y. Huang, F. Chen, S. Lv, and X. Wang, "Facial expression recognition: A survey," *Symmetry*, vol. 11, no. 10, p. 1189, Sep. 2019.
[26] H. Ge, Z. Zhu, Y. Dai, B. Wang, and X. Wu, "Facial expression recognition based on deep learning," *Comput. Methods Programs Biomed.*, vol. 215, Nov. 2022, Art. no. 106621.
[27] A. S. Vyas, H. B. Prajapati, and V. K. Dabhi, "Survey on face expression recognition using CNN," in *Proc. 5th Int. Conf. Adv. Comput. Commun. Syst. (ICACCS)*, Mar. 2019, pp. 102–106.
[28] F. Z. Canal, T. R. Müller, J. C. Matias, G. G. Scotton, A. R. de Sa Junior, E. Pozzebon, and A. C. Sobieranski, "A survey on facial emotion recognition techniques: A state-of-the-art literature review," *Inf. Sci.*, vol. 582, pp. 593–617, Jan. 2022.
[29] A. Al-Ani and M. Deriche, "An optimal feature selection technique using the concept of mutual information," in *Proc. 6th Int. Symp. Signal Process. Appl.*, 2001, pp. 477–480.
[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.
[31] 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.
[32] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1195–1215, Jul. 2022.
[33] R. K. Mishra, S. Urolagin, J. A. Arul Jothi, and P. Gaur, "Deep hybrid learning for facial expression binary classifications and predictions," *Image Vis. Comput.*, vol. 128, Dec. 2022, Art. no. 104573.
[34] Z. Song, K. Nguyen, T. Nguyen, C. Cho, and J. Gao, "Spartan face mask detection and facial recognition system," *Healthcare*, vol. 10, no. 1, p. 87, Jan. 2022.
[35] M. Aouayeb, W. Hamidouche, C. Soladie, K. Kpalma, and R. Seguier, "Learning vision transformer with squeeze and excitation for facial expression recognition," 2021, *arXiv:2107.03107*.
[36] F. Xue, Q. Wang, and G. Guo, "TransFER: Learning relation-aware facial expression representations with transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3581–3590.
[37] F. Ma, B. Sun, and S. Li, "Facial expression recognition with visual transformers and attentional selective fusion," *IEEE Trans. Affect. Comput.*, vol. 14, no. 2, pp. 1236–1248, Oct. 2021.
[38] Q. Huang, C. Huang, X. Wang, and F. Jiang, "Facial expression recognition with grid-wise attention and visual transformer," *Inf. Sci.*, vol. 580, pp. 35–54, Nov. 2021.
[39] A. Chaudhari, C. Bhatt, A. Krishna, and P. L. Mazzeo, "ViTFER: Facial emotion recognition with vision transformers," *Appl. Syst. Innov.*, vol. 5, no. 4, p. 80, Aug. 2022.
[40] S. Kim, J. Nam, and B. C. Ko, "Facial expression recognition based on squeeze vision transformer," *Sensors*, vol. 22, no. 10, p. 3729, May 2022.
[41] C.-L. Fu, Z.-C. Chen, Y.-R. Lee, and H.-y. Lee, "AdapterBias: Parameter-efficient token-dependent representation shift for adapters in NLP tasks," 2022, *arXiv:2205.00305*.
[42] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, "Training deep networks for facial expression recognition with crowd-sourced label distribution," in *Proc. 18th ACM Int. Conf. Multimodal Interact.*, Tokyo, Japan, Oct. 2016, pp. 279–283.
[43] 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*, Jun. 2010, pp. 94–101.
[44] S. Kim and H. Kim, "Deep explanation model for facial expression recognition through facial action coding unit," in *Proc. IEEE Int. Conf. Big Data Smart Comput.*, Feb. 2019, pp. 1–4.
[45] M.-I. Georgescu, R. T. Ionescu, and M. Popescu, "Local learning with deep and handcrafted features for facial expression recognition," *IEEE Access*, vol. 7, pp. 64827–64836, 2019.
[46] X. Wu, J. He, Q. Huang, C. Huang, J. Zhu, X. Huang, and H. Fujita, "FER-CHC: Facial expression recognition with cross-hierarchy contrast," *Appl. Soft Comput.*, vol. 145, Sep. 2023, Art. no. 110530.
[47] G. Meena, K. K. Mohbey, A. Indian, M. Z. Khan, and S. Kumar, "Identifying emotions from facial expressions using a deep convolutional neural network-based approach," *Multimedia Tools Appl.*, vol. 83, no. 6, pp. 15711–15732, Jul. 2023.
[48] N. Kumari and R. Bhatia, "Deep learning based efficient emotion recognition technique for facial images," *Int. J. Syst. Assurance Eng. Manage.*, vol. 14, no. 4, pp. 1421–1436, Aug. 2023.
[49] N. Kumar H N, A. S. Kumar, G. Prasad, and M. A. Shah, "Automatic facial expression recognition combining texture and shape features from prominent facial regions," *IET Image Process.*, vol. 17, no. 4, pp. 1111–1125, Mar. 2023.
[50] H. Boughanem, H. Ghazouani, and W. Barhoumi, "Multichannel convolutional neural network for human emotion recognition from in-the-wild facial expressions," *Vis. Comput.*, vol. 39, no. 11, pp. 5693–5718, Nov. 2023.

MUSTAQEEM KHAN (Member, IEEE) received the Ph.D. degree in software engineering from Sejong University, Seoul, Republic of Korea. Currently serving as a lead researcher in MBZUAI, MCR Lab, his primary research focus encompasses affective computing, computer vision, and emotion recognition. Additionally, his academic pursuits extend to areas such as audio digital signal processing, speech processing, speech synthesis, image and video processing, energy analytics, consumption predictions, and generation. With a notable presence in the academic community, he holds roles as an Associate and Guest Editor, along with serving as a professional reviewer for esteemed journals and conferences. He is a member of ACM, CTSoc, and Gold Medalist.

ABDULMOTALEB EL SADDIK (Fellow, IEEE) is currently an Enterprise Professor with MBZUAI, United Arab Emirates, while holding the esteemed position of a Distinguished University Professor with the University of Ottawa, Canada. Widely regarded as an internationally recognized scholar, he has significantly advanced the fields of intelligent multimedia computing, communications, and applications. Through his work, individuals can engage in real-time interactions with one another and their smart digital representations in the metaverse, ensuring security, and fostering seamless connectivity. His research interests include leveraging AI, the IoT, SN, AR/VR, haptics, and 5G technologies to establish digital twins that enhance citizens' quality of life. Recognized for his outstanding contributions, he has been elected as a fellow of the Royal Society of Canada, Canadian Academy of Engineering, and the Engineering Institute of Canada. He serves as the Editor-in-Chief for *ACM Transactions on Multimedia Computing, Communications, and Applications* (ACM TOMM). With a prolific academic career, he has coauthored ten books and published more than 800 research contributions. Moreover, he has chaired more than 50 conferences and workshops and mentored more than 150 researchers. His exemplary track record includes securing research grants and contracts exceeding 20 million. Notably, he authored the influential book *Haptics Technologies: Bringing Touch to Multimedia*. Additionally, he holds the title of an ACM Distinguished Scientist.

MOHAMED DERICHE (Senior Member, IEEE) received the Engineer's degree from the National Polytechnic School of Algiers and the M.Sc. and Ph.D. degrees from the University of Minnesota, USA. He joined QUT, Australia, in 1994. In 2001, he joined the King Fahd University of Petroleum and Minerals, Saudi Arabia, where he led the DSP Group. In 2021, he joined Ajman University, United Arab Emirates, as a Professor of AI/ML. He published more than 300 articles on different aspects of signal/image processing. Moreover, he has chaired more than 20 conferences and delivered more than 15 keynote/plenary talks worldwide. He supervised more than 50 graduate theses. His research interests include different aspects of multimedia signal/image processing and AI/ML applications. He was a recipient of several prestigious awards, including the Shauman Award for Best Researcher in the Arab World and the Excellence in Research Awards at both KFUPM and Ajman University.

WAIL GUEAIEB (Senior Member, IEEE) received the bachelor's and master's degrees in computer engineering and information science from Bilkent University, Turkey, in 1995 and 1997, respectively, and the Ph.D. degree in systems design engineering from the University of Waterloo, Canada, in 2001. He is currently a Professor with the School of Electrical Engineering and Computer Science (EECS), University of Ottawa, Canada. He also founded and directed the Machine Intelligence, Robotics, and Mechatronics (MIRaM) Laboratory, EECS. He is the author/coauthor of more than 120 patents and papers in highly reputed journals and conferences. His research interests include intelligent mechatronics, robotics, and applied computational intelligence. He has served as an Associate Editor, an Guest Editor, and the Program (Co-)Chair for several international journals and conferences, such as IEEE TRANSACTIONS ON INSTRUMENTATION AND MEASUREMENT, IEEE/ASME TRANSACTIONS ON MECHATRONICS, and the IEEE Conference on Decision and Control. He is an Associate Editor of *ASME Journal of Dynamic Systems, Measurement, and Control* and *International Journal of Robotics and Automation*.