2023 IEEE 19th International Conference on Body Sensor Networks (BSN) | 979-8-3503-3841-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/BSN58485.2023.10331041

Towards Efficient Deep Learning Models for Facial Expression Recognition using Transformers

Farshad Safavi
Department of Computer Science and Electrical Engineering
University of Maryland Baltimore County
Baltimore, USA
fsafavi1@umbc.edu

Kulin Patel
Department of Computer Science and Electrical Engineering
University of Maryland Baltimore County
Baltimore, USA
kulinp1@umbc.edu

Ramana Kumar Vinjamuri
Department of Computer Science and Electrical Engineering
University of Maryland Baltimore County
Baltimore, USA
rvinjam1@umbc.edu

***Abstract*—Facial expression recognition (FER) is crucial in various healthcare applications, including pain assessment, mental disorder diagnosis, and assistive robots that require close interaction with humans. While heavyweight deep learning models can achieve high accuracy for FER, their computational cost and memory consumption often need optimization for portable and mobile devices. Therefore, efficient deep learning models with high accuracy are essential to enable FER on resource-constrained platforms. This paper presents a new efficient deep-learning model for facial expression recognition. The model utilizes Mix Transformer (MiT) blocks, adopted from the SegFormer architecture, along with a supplemented fusion block. The efficient self-attention mechanism in the transformer focuses on relevant information for classifying different facial expressions while significantly improving efficiency. Furthermore, our supplemented fusion block integrates multiscale feature maps to capture both fine-grained and coarse features. Experimental results demonstrate that the proposed model significantly reduces the computational cost, latency, and the number of learnable parameters while achieving high accuracy compared with the previous state-of-the-art (SOTA) on the FER2013 and AffectNet dataset.**

Keywords— Facial Expression Recognition, Deep learning, Classification, Emotion detection, Transformer

I. INTRODUCTION

Facial expression recognition (FER) systems offer potential for numerous healthcare applications particularly in areas such as pain assessment, autism spectrum disorder diagnosis, and human-robot collaboration. Facial expressions are non-verbal cues that facilitate human communication and convey significant messages. While humans can experience complex and blended emotions, it is essential to accurately recognize fundamental expressions like anger, disgust, fear, happiness, and sadness. FER systems enable social robots to perceive and understand basic human emotions, allowing them to interact appropriately and effectively within the human space. Consequently, our study on automatic facial expression recognition (FER) systems directly applies to human-robot collaboration in medical facilities, enabling assistive robots to effectively aid patients.

Despite the availability of conventional feature extraction methods such as Local Binary Patterns (LBP) [1] and Scale Invariant Feature Transform (SIFT) [2], deep learning-based approaches in Facial Expression Recognition (FER) have gained more popularity due to their ability to automatically learn complex features from raw image data. Facial expressions rely heavily on specific facial regions, such as the eyes and mouth, which are referred to as facial landmarks. In contrast, other areas, such as the hair and jawline, have little impact on emotional expressions[3]. Consequently, facial landmark detections can yield remarkable outcomes in controlled laboratory settings. Recent studies have demonstrated the efficacy of attention mechanisms in improving the performance of convolutional neural networks for classifying facial expressions. Local (multi) Head Channel (self-attention) for facial recognition proposed a novel self-attention module that can be integrated into Convolutional Neural Networks (CNNs) [4].

Moreover, in the context of image segmentation some architectures have been shown to be effective in retaining useful pixel-level information. The Residual Masking Network [3] boost the performance of CNNs in facial expression tasks by generating segmentation masks that highlights the most informative part of the face. However, enhancing network performance through increased attention using segmentation blocks comes at a significant computational cost. These models can be inherently complex and computationally expensive, which often requires high computational and storage resources.

After conducting a thorough analysis of various real-time semantic segmentation models[5], [6], we determined that Mix Transformer (MiT) [7] Blocks are an excellent choice for use in FER systems for two reasons. Firstly, MiT blocks exhibit enhanced efficiency through the implementation of efficient self-attention, as observed in SegFormer architectures[7]. Secondly, hierarchical structure of MiT blocks can effectively extract both fine-grained and coarse features, act as a pixel-level landmark detection of the system. MiT blocks are utilized for the first time in the FER system context in this research, resulting in improved accuracy and speed.

Our main contribution in this research is the development of a novel lightweight deep learning architecture for facial expression recognition (Fig. 1). Our model leverages the power of a Mix Transformer (MiT) hierarchy along with a supplemented fusion block to achieve remarkable results. Furthermore, we conducted a thorough benchmarking analysis of our architecture against prior methods in terms of accuracy, computational complexity, latency, and number of learnable parameters on FER13 dataset[8]. To further corroborate the model's effectiveness and generalizability in

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

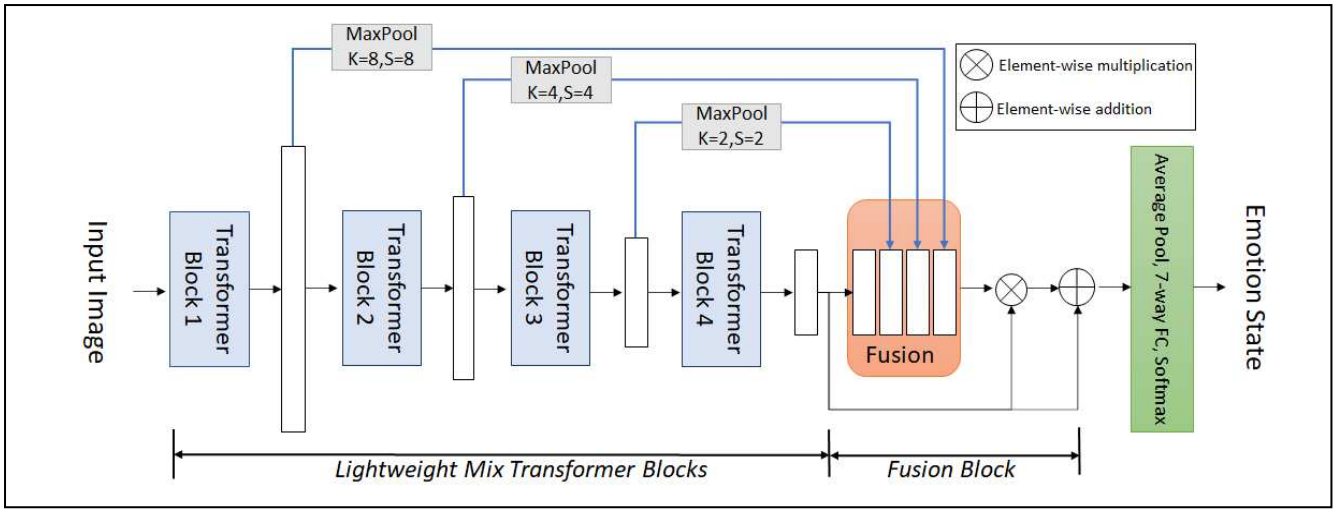

Fig. 1. The efficient architecture consists of hierarchical Mix Transformer blocks (MiT) and supplemented fusion mechanism integrating high-resolution coarse features and low-resolution fine features, which increases the network's accuracy.

real-world scenarios, we trained our model on the AffectNet dataset[9]. Our results demonstrate that our approach outperforms heavyweight methods by a significant margin in terms of efficiency, while maintaining comparable levels of accuracy. These findings have important implications for applications where low latency and computational complexity are critical factors.

II. Efficient Deep-Learning Architecture

In this section, we describe the architecture of our proposed efficient deep learning model, which aims to achieve high efficiency while maintaining accuracy for FER predictions. As depicted in Figure 2, our efficient architecture consists of hierarchical Mix Transformer blocks (MiT) [7] adopted from Transformer encoder of SegFormer architecture and additional fusion mechanism to concatenated multiscale feature maps along the channel dimension. Our supplemented fusion mechanism integrates high-resolution coarse features and low-resolution fine features which increases the network's accuracy.

An input image of size H × W × 3 initially is partitioned into patches of size 4 × 4. These patches go through Mix Transformer Blocks (MiT) which generate multi-level features. Feature maps resolutions are scaled down to 1/4, 1/8, 1/16, and 1/32 of the original image size. Then we integrate feature maps of different spatial sizes using our fusion method. Finally, the network utilizes an average pooling layer followed by a fully connected layer with SoftMax activation. This final layer produces outputs that correspond to seven distinct facial expression states. In the remaining part of this section, we provide the description of various components and addition fusion method that we incorporate into MiT blocks.

A. Mix Transformer Blocks

In this design we adopted lightweight Mix Transformer (MiT) which is encoder block of semantic segmentation framework SegFormer[7]. As we can see in TABLE I, MiT model generates four hierarchical feature maps F_i contains both coarse features and fine-grained features that enhance the performance of the network, where i ∈ {1, 2, 3, 4}. As evident in TABLE I, given an input image with size 224 × 224 × 3 produces hierarchical feature maps F_i with a resolution of $H_i \times W_i \times C_{i+1}$, where H_i, W_i ∈ {56, 28, 14, 7}, and C_{i+1} ∈ {32, 64, 128, 256} which is larger than Ci. The Transformer Blocks are composed of Overlapped Patch Merging, Efficient Self-Attention, and Mix-FFN[7]. To explain, Overlapped Patch Merging shrinks feature maps while preserving their local continuity. In addition, it allows patch of size $N \times N \times 3$ to merge into a compact $1 \times 1 \times C$ vector. Efficient Self-Attention blocks reshape and transform k, which is similar to the original multi-head self-attention process, into k', thereby compressing the computational complexity of the self-attention mechanism [7].

$$Attention(Q, K', V) = Softmax(Qk'^{T}/\sqrt{d_{head}})V \quad (1)$$

Finally, Mix-FFN uses mix of MLP and a 3×3 Conv in the feed-forward network instead of positional embedding[7].

In summary, The Mix Transformer Blocks (MiT) offer a smaller alternative to traditional vision transformers (ViT). Their efficient self-attention mechanism and exclusion of positional embedding make them highly suitable for enhancing the efficiency of Facial Expression Recognition (FER) classification without compromising accuracy.

B. Fusion Block

This block integrates the feature maps F_i to capture both low resolution fine features and high-resolution coarse features. As shown in Figure 2, we first apply max pooling operations with different kernel and stride sizes of 8, 4, and 2 to downsample feature maps F_1, F_2, and F_3. These downsampled feature maps, along with F_4, are then concatenated together and the number of channels is reduced using a 1×1 convolutional layer to produce F_R. we assume that F_R would enhance the accuracy of the feature maps, consequently, we use the output of fusion block F_R to score element-wisely the importance of the final output of Transformer blocks F4 using following operation:

$$F_N = F_4 + F_4 \otimes F_R \quad (2)$$

Finally, F_N passes through the average pooling and fully

TABLE I
EFFICIENT MIX TRANSFORMER WITH FUSION BLOCK

Layer name	Output size	Detail
Transformer Block 1	32 × 56 × 56	K = 7, S = 4, P = 3
Transformer Block 2	64 × 28 × 28	K = 3, S = 2, P = 1
Transformer Block 3	128 × 14 × 14	K = 3, S = 2, P = 1
Transformer Block 4	256 × 7 × 7	K = 3, S = 2, P = 1
Fusion Block	256 × 7 × 7	MaxPool2, Concat
Average pooling	256 × 1 × 1	
FC, Softmax	7	

connected layer with a dropout layer (dropout rate of 0.4) and a linear layer that maps the input to 7 output classes. The detailed kernel sizes, strides and paddings for all layers are provided in TABLE I.

III. EXPERIMENT

A. FER2013 Dataset

In this research, we used FER2013 (Facial Expression Recognition)[8] as a benchmark dataset for comparing the accuracy of different FER models. The dataset comprises 33,572 facial images with resolutions of 48x48 pixels in grayscale. The dataset was categorized into seven standard classes, with distributions of Angry (4,953), Disgust (547), Fear (5,121), Happy (8,989), Sad (6,077), Surprise (4,002), and Neutral (6,198). FER2013 achieves human-level accuracy of 65±5% and the top algorithm attains 76.82% accuracy in labeling facial expressions[3].

B. AffectNet Dataset

To validate the model's effectiveness and generalizability in real-world scenarios, we trained the proposed model on the AffectNet dataset, a large FER dataset in the wild which includes manually labeled images in eight emotional states (neutral, happy, angry, sad, fear, surprise, disgust, contempt). Our training set consisted of 287,657 images from AffectNet dataset, and testing was performed on the official test set of 4,000 images (500 per emotional class).

C. Experimental Setup

The training images are resized to 224 × 224 and transformed to the RGB format to match the requirements of pre-trained models in ImageNet. Data augmentation techniques are employed during training to prevent overfitting. These techniques encompass left-right flipping and rotation within the range of [-30, 30] degrees. However, other data augmentation techniques did not improve the model's performance. Each experiment is conducted for a maximum of 50 epochs and will terminate if the validation accuracy does not improve for more than eight consecutive steps. A batch size of 48 is employed with an initial learning rate of 0.0001, which is reduced by a factor of ten if the validation accuracy does not improve for two consecutive epochs. The momentum is set to 0.9 and weight decay to 0.001. We train all models on a single system with NVIDIA GeForce RTX 3090 GPU with an Intel Core i9 processor. Experiments using different networks are conducted under the same settings.

D. Evaluation Matrices

In this section, we will outline the evaluation metrics employed to assess the accuracy and efficiency of the networks in this experiment.

To evaluate the accuracy of classification models, we use following:

$$Accuaracy = (TP + TN)/\,(TP + TN + FP + FN) \quad (3)$$

The correctly predicted pixels are referred to as true positives (TP), while those correctly identified as not belonging to a specific class are known as true negatives (TN). Pixels that belong to the category but are incorrectly predicted as a different type are called false negatives (FN). Lastly, the pixels mistakenly indicated as belonging to the class are referred to as false positives (FP).

To measure the efficiency of our models, we used three metrics: the number of trainable parameters, computational complexity (Flops), and inference time (TABLE II).

1) Learnable Parameters

This metric quantifies the extent of model complexity by measuring the total number of learnable parameters in a feed-forward neural network.

2) Flops

The Flops measures computational complexity based on the number of floating-point operations per second.

3) Inference Time

The inference time is calculated on a single RTX 3090 GPU using CUDA 11.7, and PyTorch 1.13.0. After initializing the GPU with dummy examples, the network is run 300 times with an input resolution of 224 × 224 and a batch size of 48. The average time is then reported. For real-time model consideration, the standard video streaming rate is 24 frames per second (fps), meaning that if a model processes an image in less than 41ms, it can be categorized as a real-time model.

IV. RESULTS AND ANALYSIS

To evaluate the method's performance on the FER2013 public dataset, we categorized the network classifications into two distinct group. First, we focused on networks that yield high accuracy on FER2013, including ResmaskingNet [3], Resnet151 [10], Densenet121 [11], RessAttNet56 [12], Cbam_resnet50 [13]. Second, we analyzed recent efficient networks, namely MobileNetV3 [14], MobiExpressNet [15], Improved MobileNetV3(imp-MobileNetV3) [16] and RASN [17]. As depicted in TABLE II, among the assessed methods in the first group on the FER2013 dataset, our approach displays the highest efficiency across all efficiency metrics: Flops, the number of learnable parameters, and inference time. Furthermore, within the lightweight category located in the lower section of TABLE II, our model attains the highest accuracy among all models, while demonstrating comparable efficiency.

Our approach achieves a low Flops value of 425 million through efficient transformer block utilization via compact hierarchical design and exclusion of positional embeddings from the original vision transformer structure. This yields high accuracy with minimal computational load, comparable to lightweight methods like MonileNetV3 [14]. Similarly, the proposed deep network excels the efficiency with only 3.45

TABLE II
PERFORMANCE EVALUATION OF NETWORKS ON FER2013

Light Models	Flops	Parm	Time	Acc
ResAttNet56	6.33B	29.77M	17.69ns	72.63%
Densenet121	2.89B	6.96M	19.57ns	73.16%
Resnet152	11.60B	58.16M	23.66ns	73.22%
Cbam_resnet50	4.14B	26.05M	16.94ns	73.39%
ResMaskingNet	26.76B	142.9M	17.63ns	74.14%
MobileNetV3	0.23B	5.48M	7.48ns	64.78%
MobiExpressNet	1.09M	75.08K	-	67.96%
Imp-MobileNetV3	0.19B	1.29M	8.93ns	68.14%
RASN	1.83B	14.4M	-	71.44%
Proposed Method	**0.425B**	**3.45M**	**7.74 ns**	**73.47%**

TABLE II. Evaluating the Efficiency and Accuracy of All Models. The results of the proposed method are emphasized in bold, presenting Floating-Point Operations per Second (Flops), Learnable Parameters (Parms), Inference Time (Time), and Accuracy (Acc).

million parameters which is 41 times lower than that of ResMaskingNet. This efficiency is exceeds lightweight models like RASN[17]. By employing a compact model with fewer parameters, our proposed method reduces memory requirements and computational overhead. Furthermore, our model showcases impressive speed in terms of inference, as evidenced by an average inference time of 7.74 nanoseconds (ns). This characteristic highlights its suitability for real-time applications, on par with lightweight approaches like MobileNetV3 and Imp- MobileNetV3[16].

In addition to its efficiency, the proposed network achieves a high accuracy of 73.47%, which is the second-best result obtained in the evaluation table (TABLE II). To further validate the model's efficacy and its ability to generalize in real-world scenarios, we conducted training on AffectNet [9]. This yielded a 60.9% accuracy, which is comparable to 63.05% accuracy achieved by the state-of-the-art (SOTA) performance [18] on the AffectNet dataset. The higher accuracy is due to our hierarchical Transformer Blocks with a larger effective receptive field (ERF) than traditional CNN layers, showcasing a balance between efficiency and accuracy. As seen in TABLE II, our model stands out for efficiency in terms of Flops, number of parameters, and inference, compared to networks in the first group that achieve high accuracy on FER2013. Furthermore, unlike the second group which compromises accuracy for efficiency, our model maintains a high level of accuracy on FER2013 and AffectNet datasets.

V. CONCLUSION

This paper introduces a novel efficient deep learning model for facial expression recognition (FER) using a Mix Transformer (MiT) hierarchy and a supplemented fusion block. The model incorporates a self-attention mechanism to enhance attention at the pixel-level landmark segments. In addition, the fusion block combines high-resolution coarse features with low-resolution fine features to achieve improved accuracy. Experimental results demonstrate that the proposed method achieves high efficiency while maintaining competitive accuracy on the FER2013 and AffectNet dataset. Future work includes applying this method to develop a human-robot collaboration system, where facial emotion recognition, combined with bio-signals, collaborates with a robot.

REFERENCES

[1] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, 2002, doi: 10.1109/TPAMI.2002.1017623.

[2] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004, doi: 10.1023/B:VISI.0000029664.99615.94.

[3] L. Pham, T. H. Vu, and T. A. Tran, "Facial Expression Recognition Using Residual Masking Network," in *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 4513–4519. doi: 10.1109/ICPR48806.2021.9411919.

[4] R. Pecoraro, V. Basile, V. Bono, and S. Gallo, "Local Multi-Head Channel Self-Attention for Facial Expression Recognition," *CoRR*, vol. abs/2111.0, 2021, [Online]. Available: https://arxiv.org/abs/2111.07224

[5] F. Safavi and M. Rahnemoonfar, "Comparative Study of Real-Time Semantic Segmentation Networks in Aerial Images During Flooding Events," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 16, pp. 15–31, 2023, doi: 10.1109/JSTARS.2022.3219724.

[6] F. Safavi, T. Chowdhury, and M. Rahnemoonfar, "Comparative Study Between Real-Time and Non-Real-Time Segmentation Models on Flooding Events," in *2021 IEEE International Conference on Big Data (Big Data)*, 2021, pp. 4199–4207. doi: 10.1109/BigData52589.2021.9671314.

[7] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers," in *Advances in Neural Information Processing Systems*, 2021, vol. 34, pp. 12077–12090. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2021/file/64f1f27bf1b4ec22924fd0acb550c235-Paper.pdf

[8] I. J. Goodfellow *et al.*, "Challenges in Representation Learning: A Report on Three Machine Learning Contests," in *Neural Information Processing*, 2013, pp. 117–124.

[9] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "{AffectNet}: A Database for Facial Expression, Valence, and Arousal Computing in the Wild," *{IEEE} Trans. Affect. Comput.*, vol. 10, no. 1, pp. 18–31, Jan. 2019, doi: 10.1109/taffc.2017.2740923.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[11] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2261–2269. doi: 10.1109/CVPR.2017.243.

[12] F. Wang *et al.*, "Residual Attention Network for Image Classification," *CoRR*, vol. abs/1704.0, 2017, [Online]. Available: http://arxiv.org/abs/1704.06904

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

[14] A. Howard *et al.*, "Searching for MobileNetV3," 2019, pp. 1314–1324. doi: 10.1109/ICCV.2019.00140.

[15] S. F. Cotter, "MobiExpressNet: A Deep Learning Network for Face Expression Recognition on Smart Phones," in *2020 IEEE International Conference on Consumer Electronics (ICCE)*, 2020, pp. 1–4. doi: 10.1109/ICCE46568.2020.9042973.

[16] X. Liang, J. Liang, T. Yin, and X. Tang, "A lightweight method for face expression recognition based on improved MobileNetV3," *IET Image Process.*, vol. 17, no. 8, pp. 2375–2384, 2023, doi: https://doi.org/10.1049/ipr2.12798.

[17] J. Yang, Z. Lv, K. Kuang, S. Yang, L. Xiao, and Q. Tang, "RASN: Using Attention and Sharing Affinity Features to Address Sample Imbalance in Facial Expression Recognition," *IEEE Access*, vol. 10, pp. 103264–103274, 2022, doi: 10.1109/ACCESS.2022.3210109.

[18] A. V Savchenko, L. V Savchenko, and I. Makarov, "Classifying emotions and engagement in online learning based on a single facial expression recognition neural network," *IEEE Trans. Affect. Comput.*, pp. 1–12, 2022, doi: 10.1109/TAFFC.2022.3188390.