EmoPAtt-Lite: Lightweight Facial Emotion Recognition

Ismail Ben seddik, *Luddy School of Informatics, Computing and Engineering, Indiana University*
Adebowale Emmanuel Adelekan, *Luddy School of Informatics, Computing and Engineering, Indiana University*

***Abstract*—Facial expressions are a fundamental component of human communication, conveying a wide range of emotions. However, automatic facial expression recognition (FER) in the wild remains challenging, particularly under adverse conditions. Recent advances in computer vision have demonstrated the effectiveness of deep neural networks for FER, but their deployment is often constrained by the need for substantial computational resources. To address this, we propose EmoPAtt-Lite, a lightweight architecture for facial expression recognition that combines a truncated MobileNetV1 backbone with a novel attention-based classifier, augmented by a Spatial Transformer Network (STN) for spatial adaptability and a Squeeze-and-Excitation (SE) block for enhanced feature recalibration. Despite its compact size of only 1.3M parameters, EmoPAtt-Lite achieves state-of-the-art performance on the FER2013 benchmark, reaching an accuracy of 79.35%, thus demonstrating that high recognition accuracy can be attained without heavy computational demands.**

***Index Terms*—Facial expression recognition, MobileNetV1, Attention-based classifier, Emotion classification.**

I. Introduction

THE intricacies of facial expression present a captivating dimension of nonverbal human communication, involving a spectrum of facial muscle movements capable of conveying diverse emotions and mental states. This form of expression holds significant weight in interpersonal communication, as it constitutes a powerful channel for transmitting information, surpassing the mere spoken word. While words contribute a mere 7% to effective communication, voice tone and body language account for 38% and 55% respectively [1]. Given the pivotal role of facial expressions, there has been a burgeoning interest in automated Facial Expression Recognition (FER) technology. The potential applications of FER are vast, spanning multiple domains such as education, healthcare and human-machine interaction. For instance, in educational settings, FER could offer insights into teaching effectiveness and quality [2] [3], while in healthcare, it could aid in the psychological assessment of patients [4] [5].

Unlike other image classification tasks, FER presents distinct challenges stemming from both inter-class similarities and intra-class differences in human facial expressions [6]. Inter-class similarities pertain to the subtle distinctions between facial expressions, complicating the differentiation of nuanced variations and accurate recognition. Conversely, intra-class differences refer to the diversity within FER databases, where images of a single expression may encompass various subjects with differing facial structures, genders, ages, and races. This variability can impede learning efficacy, as models may struggle to generalize across subjects, resulting in diminished accuracy and reliability. For instance, distinguishing between an angry and a disgusted expression may prove challenging due to minimal visual disparities, while variations between individuals within the same expression category can be substantial. Moreover, existing studies exposed FER challenges in "in-the-wild" databases, notably the recognition of negative expressions, FER under adverse conditions, and the reliance on large neural networks. The limited availability of negative expression images online hampers the creation of representative datasets reflecting real-world scenarios, potentially leading to class imbalances favoring positive expressions. FER under challenging conditions involves recognizing facial expressions when subjects are posed at certain angles or when facial features are partially obscured by objects. Accurately identifying these instances is crucial, particularly as they mirror the conditions often encountered in practical applications. Furthermore, as efforts intensify to enhance classification performance, there's a growing tendency towards employing large neural networks. However, considering the computing constraints of downstream applications, FER methodologies should be accessible without requiring extensive computational resources.

This paper introduces EmoPAtt-Lite, a lightweight network relying on MobileNetV1. We employ a modified version of MobileNetV1, pre-trained on ImageNet, as the foundational feature extractor [6]. Accompanying this is a novel attention classifier designed to enhance learning from the feature maps produced by the lightweight extractor. Additionally, we incorporate a Spatial Transformer Network (STN) to enable neural networks to learn and apply spatial transformations to input data. Moreover, a Squeeze-and-Excitation (SE) block is included to enhance the representational capacity of the patch extraction network [7]. Our main contribution lies in successfully incorporating a lightweight patch extraction block into the truncated MobileNetV1 architecture, achieving state-of-the-art performance on the FER2013 dataset. This was achieved with a network boasting very few parameters and minimal training requirements, thus mitigating costs and complexities associated with traditional deep learning models [8].

II. Related Work

With the notable advancements and increasing prominence of deep learning in computer vision, particularly in tasks like

This work is conducted under the auspices of the Computer Vision class, taught by Dr. David Crandall, within the Department of Computer Science at the Luddy School of Informatics, Computing, and Engineering, Indiana University.

image classification, numerous studies have proposed various deep learning methodologies to tackle automated Facial Expression Recognition (FER) on the FER2013 dataset. The primary aim across these studies has been to attain optimal classification accuracy.

Prior to the dominance of deep learning in computer vision, researchers explored content-based image retrieval (CBIR) methods that relied heavily on handcrafted features such as color, texture, and object correlation. For example, Muhammad Nawaz et al. proposed an object identification system where images were indexed using both low-level features (color histograms, textures) and high-level descriptors (geometrical object shapes and their correlations) [9]. Their framework integrated a knowledge base to capture relationships among objects, thereby improving retrieval accuracy. However, such systems faced limitations in bridging the semantic gap—the disconnect between machine-extracted low-level features and the high-level concepts humans rely on for interpretation.

This gap has driven the shift toward deep learning-based methods, where Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) directly learn hierarchical feature representations for tasks like Facial Expression Recognition (FER). Unlike CBIR approaches, modern FER methods emphasize learning discriminative representations of facial regions to capture subtle emotional cues, thereby reducing reliance on handcrafted descriptors.

Various deep learning methodologies have been investigated for Facial Expression Recognition (FER), but Convolutional Neural Networks (CNNs) have consistently emerged as the favored approach in existing literature. Numerous CNN-based methods, as referenced in [10], [11] and [12], aim to improve FER accuracy by harnessing local information through additional modules. The advancement of CNN architectures has introduced several innovations, including residual connections [13], bottleneck design [14], and depth-wise separable convolutions [15], among others. However, architectures incorporating these innovations often become complex and feature a higher number of training parameters, necessitating larger datasets and longer training times. One notable example is the Residual Masking Network, which presents a novel Masking Idea to enhance CNN performance in facial expression tasks [16]. It employs a segmentation network to refine feature maps, enabling the network to concentrate on relevant information for accurate decisions. It combines the widely-used Deep Residual Network with a U-net-like architecture to form the Residual Masking Network. Although this ensemble learning model achieves state-of-the-art accuracy on the FER2013 dataset, it comprises over 142M parameters and demands considerable time and computational resources for training.

The Vision Transformer (ViT) architecture presents a novel adaptation of the Transformer architecture for computer vision tasks. It segments images into smaller, non-overlapping patches and transforms them into 1D sequences before processing them as a sequence using a Transformer model. Recently, researchers have begun exploring the application of Transformers or the ViT architecture in Facial Expression Recognition (FER), as evidenced by studies referenced in [17], [18], and [19]. Motivated by their performance in various tasks, the adoption of vision transformers for FER has shown promise. However, these architectures often entail a larger number of parameters compared to CNN-based methods, despite their notable performance. Nonetheless, the remarkable performance of the ViT architecture has prompted interest in integrating its principles into other architectures. For instance, [6] drew inspiration from ViT to enhance FER performance within the MobileNetV1 backbone. Although the patch extraction block is influenced by ViT, there are differences in implementation details. Firstly, the design and placement of the patch extraction block vary. In ViT, the patch extraction mechanism is a single-layer convolution positioned at the start of the architecture, whereas in the proposed PAtt-Lite, it is a multi-layer convolution placed within the architecture. This placement enables the proposed method to fully leverage the pretrained weights of the backbone MobileNetV1, trained on ImageNet samples of size 224×224.

Other research has shown effectiveness and impressive accuracy, albeit with the drawback of increased complexity leading to costly training. A study by Vignesh et al. introduced a novel deep learning model called Segmentation VGG-19 [20]. This model enhances Facial Expression Recognition (FER) by incorporating segmentation blocks based on U-Net into the VGG-19 architecture. By integrating these segmentation blocks between VGG-19 layers, the model effectively emphasizes significant features in the feature map, thereby enhancing the feature extraction process. On a different note, EmoNeXt has demonstrated its superiority over existing state-of-the-art deep learning models in terms of emotion classification accuracy on the FER2013 dataset [7]. It presents a novel deep learning framework for facial expression recognition, based on an adapted ConvNeXt architecture network. EmoNeXt integrates a Spatial Transformer Network (STN) to focus on feature-rich facial regions and incorporates Squeeze-and-Excitation blocks to capture channel-wise dependencies. Furthermore, it introduces a self-attention regularization term, encouraging the model to generate compact feature vectors.

While much of the research in emotion recognition has concentrated on visual modalities, parallel efforts have been made in text-based emotion identification. For instance, recent work by Imad Rida et al. introduced a novel multilingual BERT-based framework for emotion identification from resource-constrained languages such as Roman Urdu [21]. Their study addressed the scarcity of datasets and the inherent complexity of Roman Urdu by developing the RUDE dataset, incorporating six Ekman-based emotions, and introducing a novel preprocessing tool for language standardization. Leveraging transfer learning with mBERT, their approach achieved state-of-the-art performance with an average accuracy of 91%, surpassing conventional machine learning and deep learning baselines.

Although this study is text-oriented, its emphasis on transfer learning from pre-trained models and adapting architectures for resource-constrained settings resonates with recent trends in vision-based FER, where lightweight backbones and attention mechanisms are integrated to balance performance and efficiency. Together, these works highlight the growing role of domain adaptation, pretraining, and modality-specific innova-

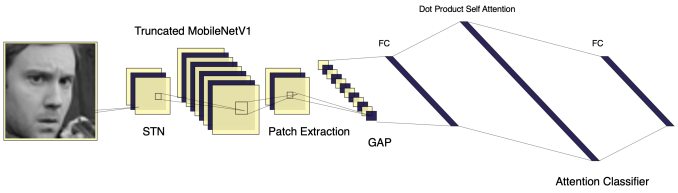

Fig. 1. Architechture of the proposed EmoPAtt-Lite.

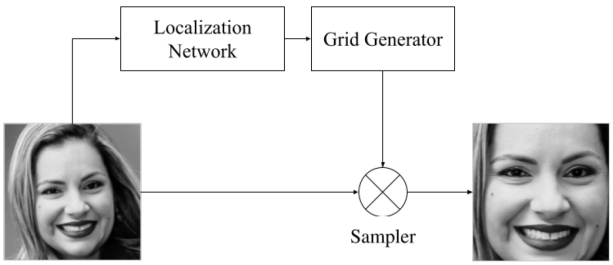

Fig. 2. The architecture of the spatial transformer network (STN).

tions in advancing the broader field of emotion recognition.

In addition to vision- and text-based modalities, related studies have also explored the intersection of psychological traits, behavioral responses, and deep learning methods. For instance, a 2022 study investigated the relationship between the Big Five personality traits and susceptibility to phishing attacks, a prevalent form of social engineering [22]. Due to the absence of publicly available datasets combining personality profiles and phishing responses, they employed a conditional Generative Adversarial Network (C-GAN) for both synthetic data generation and classification. Their findings indicate that certain personality traits significantly correlate with vulnerability to phishing, providing insights into how user psychology can influence susceptibility to cyber threats.

Although this study is situated in the domain of cybersecurity, its focus on leveraging deep learning to model human behavior, psychological attributes, and limited-data scenarios parallels ongoing challenges in FER research. Together with text-based and vision-based approaches, it highlights a broader trend in emotion and behavior recognition research: adapting advanced architectures such as BERT, CNNs, Transformers, and GANs to tackle both modality-specific constraints and the complex interplay between human psychology and observable signals.

III. Methods

In this section, we present the methodology employed to construct a comprehensive deep learning model tailored specifically for facial emotion recognition. Illustrated in Fig. 1, the architecture of our proposed EmoPAtt-Lite comprises a Spatial Transformer Network (STN) followed by a truncated MobileNetV1. Leveraging the pre-trained feature-extracting capabilities of MobileNetV1, the model effectively captures lower-level details from input images. The output feature maps then undergo patch extraction to isolate significant local features. By integrating a Squeeze-and-Excitation (SE) block within the patch extraction block, the model gains the capability to dynamically adjust the importance of different features within each channel. Subsequently, the attention classifier utilizes globally average pooled feature maps as input and generates probabilities corresponding to facial expressions.

A. Spatial Transformer Network

Spatial Transformer Networks (STN), as utilized in EmoNeXt, extend the concept of differentiable attention to encompass diverse spatial transformations [7]. This integration introduces a differentiable geometric transformation module into the network architecture, enabling neural networks to learn and apply spatial transformations to input data. Such capability is particularly advantageous in Facial Expression Recognition (FER), where variations in scale, rotation, and translation significantly impact performance. The spatial transformer mechanism comprises three distinct components [23], as depicted in Fig. 2 [7]. Sequentially, a localization network receives the input feature map and, through a series of hidden convolutional layers, generates the parameters of the spatial transformation to be applied to the feature map. This results in a transformation that is contingent upon the input. Subsequently, the predicted transformation parameters are utilized to generate a sampling grid by the grid generator. This grid consists of points indicating where the input map should be sampled to generate the transformed output. Lastly, the feature map and the sampling grid are inputted into the sampler, which samples the input image at the grid points to produce the final output image.

A notable advantage of STN is their capacity to autonomously learn spatial transformations as part of the neural network training process. Consequently, in EmoPAtt-Lite, an STN is integrated immediately after the input to enhance the network's capability to capture spatial dependencies, align facial features, and concentrate on pertinent regions of the input images.

B. MobileNetV1

MobileNets represent a category of efficient models tailored for mobile and embedded tasks [15]. These models are constructed on a simplified architecture that employs depthwise separable convolutions to construct lightweight deep neural networks. Through the utilization of depthwise separable convolutions, MobileNetV1 achieves a notable reduction in the quantity of model parameters and the computational complexity measured by the number of multiplication and addition operations required for inference (Mult-Adds) [6].

Depthwise separable convolution (Fig. 3) diverges from the standard convolutional operation by dividing it into two distinct operations: depthwise convolutions followed by pointwise convolutions. In depthwise convolution, a single convolutional filter is applied to each input channel, unlike conventional convolution where filters are as deep as the input. Meanwhile,

pointwise convolution can be executed using the standard convolutional operation by setting the kernel size to 1. Essentially, pointwise convolutions facilitate the blending of input channels, akin to conventional convolutions.

MobileNetV1, pre-trained on ImageNet and chosen for its lightweight CNN architecture, serves as the foundational feature extractor in our approach. This selection is made due to its capacity for straightforward fine-tuning of pre-trained weights on our benchmark datasets, minimizing the risk of overfitting to the training samples.

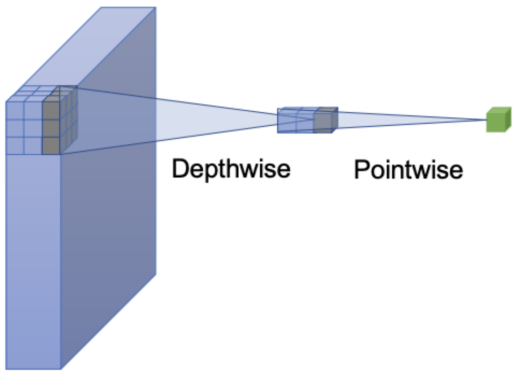

Fig. 3. Depthwise separable convolution.

C. Patch Extraction

The incorporation of the patch extraction block enhances adaptation to Facial Expression Recognition (FER) as opposed to merely fine-tuning the final layers of the pre-trained convolutional network [6]. This adjustment to the feature extractor also reduces the training duration, as a higher learning rate can be applied, unlike when fine-tuning pre-trained weights.

The patch extraction block comprises three distinct convolutional layers: the initial two layers utilize depthwise separable convolutions, followed by a pointwise convolutional layer. Subsequently, a Squeeze-and-Excitation (SE) module is employed to enable the network to dynamically recalibrate channel-wise features, thereby augmenting the representational capabilities of the feature maps extracted by the CNN.

Operating on feature maps resized to 16×16 from MobileNetV1, the initial separable convolutional layer divides the feature maps into four patches while simultaneously learning higher-level features from its input. Following this, the subsequent separable convolutional layer and the pointwise convolutional layer focus on learning higher-level features from these patched feature maps, resulting in output with dimensions of 2×2. In contrast to the standard convolutional layer typically used in conventional CNNs, the PAtt-Lite model opts for the depthwise separable convolutional layer. This selection enhances classification performance while simultaneously reducing the number of model parameters.

The Squeeze-and-Excitation (SE) method, as described in [7], is a powerful technique employed in deep learning architectures to bolster the expressive capability of CNN models. It introduces a mechanism allowing the network to dynamically adjust features at the channel level, thereby augmenting the model's ability to distinguish between classes. As depicted in Fig. 4, the SE block consists of two key operations: squeezing and exciting. In the squeezing step, global average pooling is utilized on each channel of the feature map (W, H, C), reducing its spatial dimensions to a single value per channel (1, 1, C). Following this, in the exciting phase, the squeezed values undergo transformation through a small set of fully connected layers. These layers learn specific weights for each channel, capturing the relationships among feature channels. The resulting attention weights are then element-wise multiplied with the original feature map, amplifying informative channels while diminishing less pertinent ones.

Incorporating the SE block into patch extraction enables the model to autonomously regulate the importance of various features within the extracted patches. This adaptability assists in highlighting discriminative facial features pertinent to specific expressions while mitigating irrelevant or noisy features, thereby facilitating more effective representation learning. By prioritizing the most informative channels within the feature maps, the SE block augments the model's capacity to discern subtle disparities between similar expressions, resulting in enhanced recognition accuracy. While the SE block introduces additional parameters and computations, its inclusion typically incurs a relatively modest computational overhead compared to other intricate architectures. Consequently, integrating the SE block facilitates efficient enhancement of feature representation without significantly escalating the computational burden.

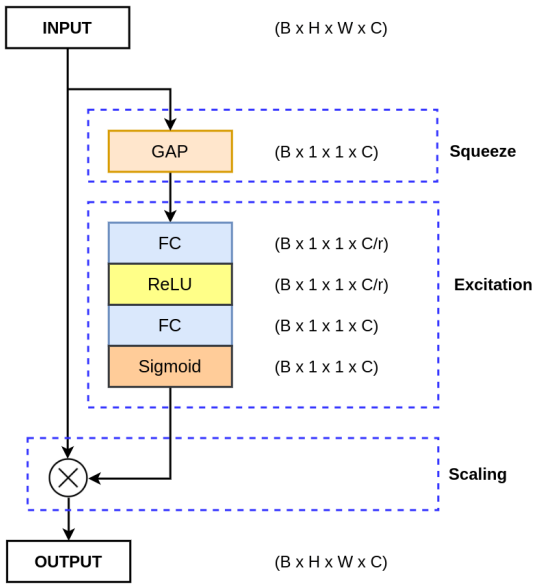

Fig. 4. The architecture of the Squeeze-and-Excitation module.

D. Attention Classifier

Self-attention serves as an attention mechanism that connects various positions within a single sequence to generate a

representation of the sequence [24]. Dot product attention, on the other hand, is a particular form of self-attention mechanism where attention weights are calculated through a dot product between the query vector and the key vector, divided by the square root of the dimension of the key vectors.

Given Q, K, and Vas the query, key, and value vectors, respectively, where the dimension of Q is equal to that of K, the dot-product self-attention score can be calculated as follows:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (1)$$

$\sqrt{d_k}$ is the dimension of the key vector k and query vector q.

An attention classifier is employed to enhance the learning of representations from both the backbone MobileNetV1 and the patch extraction block [6]. This attention classifier consists of a dot-product self-attention layer positioned between two fully connected layers within the newly integrated classifier.

IV. Results

A. Dataset

The experiments were carried out using the FER2013 dataset, which was initially introduced in the ICML 2013 Challenges in Representation Learning [8]. This dataset comprises 35,887 grayscale images, each sized at 48×48 pixels. It is partitioned into three subsets: 28,709 images for training, 3,589 images for validation, and 3,589 images for testing. The faces in the dataset are categorized into one of seven classes. Due to its substantial sample size, this dataset is widely employed in Facial Expression Recognition (FER) tasks. However, it is worth noting that the class distribution within this dataset is highly imbalanced, which presents a challenge for any deep learning model. Fig. 5 displays a selection of example data extracted from the FER2013 dataset.

Fig. 5. Sample images from FER2013.

B. Model Training

Initially, all sample images undergo a resizing operation to standardize their dimensions to 224×224. Data augmentation techniques, including random horizontal flip and random contrast adjustments, are applied. Resizing from the original 48×48 resolution presents the challenge of potentially amplifying noise and reducing the visibility of fine-grained facial cues, which are important for accurate emotion recognition. Similarly, while data augmentation improves model generalization, it also introduces the challenge of carefully balancing transformations so as not to distort subtle expressions or worsen the dataset's inherent class imbalance. The training process follows a two-stage training-finetuning approach, as outlined in [6]. During training, the pre-trained weights are kept fixed to solely adapt the new components to the output from MobileNetV1. To maximize the feature extraction performance of MobileNetV1, 59 layers are unfrozen. The sparse categorical cross-entropy loss function is utilized, with Adam serving as the optimizer as well as a batch size of 32 throughout all experiments. To enhance the stability of the proposed method, global gradient norm clipping is implemented during experiments. The initial learning rate is set to 1×10^{-3} for the initial training stage. Learning rate adjustments are made by decreasing it when the model accuracy stagnates beyond a specified patience threshold. For the finetuning process, the learning rate follows an inverse time decay schedule with an initial learning rate of 1×10^{-5}. The number of epochs for both the initial training and fine-tuning phases is determined by the early stopping callback, with restoration to the best weights upon termination of the training process. Due to computational constraints, training epochs are limited to 30, while fine-tuning epochs are set to 100. Nonetheless, increasing the number of epochs has the potential to further improve accuracy. In particular, extending training to 100 epochs and fine-tuning to 500 epochs has been shown to yield better classification results, as reported in [6]. Additionally, experiments include testing a cosine decay scheduler, albeit without yielding superior results. Different model architectures are explored, including the integration of a Convolutional Block Attention Module (CBAM) to enhance facial emotion discrimination and recognition. However, none of these attempts resulted in significant improvements, leading to the adoption of the architecture detailed in the preceding sections.

C. Results

The findings outlined in Table I underscore the superior performance of our proposed model, EmoPAtt-Lite, in comparison to existing state-of-the-art architectures trained on the FER2013 dataset. This success can be attributed to the design of EmoPAtt-Lite, which draws inspiration from EmoNeXt [7] and PAtt-Lite [6], effectively capturing and highlighting pertinent facial features for precise emotion classification. EmoPAtt-Lite achieves an accuracy of 79.35%, surpassing the current best state-of-the-art accuracy attained by EmoNeXt-XLarge (76.12%) and the accuracy achieved by an ensemble of networks as seen in Ensemble ResMaskingNet (76.82%) [16]. Additionally, within the limited number of epochs utilized for training and fine-tuning, EmoPAtt-Lite also surpasses the accuracy of plain PAtt-Lite (77.41%). This remarkable performance firmly establishes EmoPAtt-Lite as an exceptionally effective model for image classification tasks, specifically for facial emotion recognition (FER). Notably, our network boasts significantly fewer parameters and is remarkably lightweight compared to existing approaches. While EmoNeXt-XLarge achieves the highest performance for a model not utilizing ensemble learning on FER2013, necessitating over 31M parameters to achieve an accuracy of 76.12%, our model contains only around 1.3M parameters. Furthermore, with a training duration of 30 epochs followed by 100 epochs

for fine-tuning, our model achieves an outstanding accuracy of 79.35%, showcasing its efficiency and efficacy in facial emotion recognition tasks.

TABLE I
COMPARISON OF PERFORMANCE ON THE FER2013 TEST SET

Model	Accuracy (%)
GoogleNet	65.20
ConvNeXt-Large	73.46
Residual Masking Network	74.14
Segmentation VGG-19	75.97
EmoNeXt-XLarge	76.12
Ensemble ResMaskingNet	76.82
EmoPAtt-Lite	**79.35**

V. DISCUSSION

In the pursuit of facial emotion recognition, our original goal evolved into the development of EmoPAtt-Lite, a model that not only achieved remarkable accuracy (79.35%) on the FER2013 dataset but also surpassed the state-of-the-art performance set by Ensemble ResMaskingNet [16]. Our research initially aimed to create a model capable of achieving higher accuracy than existing state-of-the-art approaches while addressing concerns regarding complexity and computational resources. Through extensive experimentation, we observed that models like Ensemble ResMaskingNet, EmoNeXt, and Segmentation VGG-19, while effective, posed challenges due to their large size and complexity, requiring significant time and computational resources for training. Given our focus on real-time facial expression recognition for human-computer interaction applications, we shifted our objective towards developing a smaller, more efficient architecture.

PAtt-Lite emerged as a promising candidate due to its ability to achieve high accuracy with a lightweight architecture based on a truncated MobileNetV1 pre-trained on ImageNet [6]. Building upon this foundation, our experimentation focused on integrating key modules from EmoNeXt [7], specifically the Squeeze-and-Excitation (SE) Block and a Spatial Transformer Network (STN), into the PAtt-lite architecture. The results from our ablation study, as depicted in Table II, highlight the effectiveness of this approach. We observed that incorporating these prominent modules from EmoNeXt into the PAtt-lite architecture led to improved performance without introducing additional complexity or computational costs.

Our findings underscore the importance of balancing accuracy with efficiency, particularly in applications requiring real-time facial expression recognition. By leveraging lightweight architectures and integrating key modules from state-of-the-art models, we demonstrate the feasibility of achieving high accuracy while minimizing complexity and computational overhead. However, it is essential to acknowledge certain limitations of our study. While EmoPAtt-Lite shows promise in achieving state-of-the-art accuracy, further evaluation on diverse datasets and real-world scenarios is necessary to validate its robustness and generalizability. Additionally, while we focused on reducing complexity and computational costs, future research could explore optimizations to further enhance efficiency without compromising accuracy.

TABLE II
ABLATION STUDY FOR EMOPATT-LITE ON FER2013

STN	SE Block	Accuracy (%)
–	–	77.41
✓	–	77.63
–	✓	78.68
✓	✓	**79.35**

VI. CONCLUSION

This work introduces EmoPAtt-Lite, a deep learning framework designed for facial expression recognition. Leveraging the lightweight MobileNetV1 architecture, EmoPAtt-Lite aims to enhance classification accuracy on the FER2013 dataset. Our proposed EmoPAtt-Lite achieves state-of-the-art performance on FER2013 while maintaining a significantly reduced parameter count, totaling just 1.3M parameters. By incorporating a Spatial Transformer Network (STN) and a Squeeze-and-Excitation block within the patch extraction block, our model effectively captures intricate facial features, leading to improved emotion classification accuracy.

Our findings suggest that with adequate computational resources, EmoPAtt-Lite could potentially achieve even higher accuracy, as indicated by the outstanding 92.50% accuracy reported in [6]. Nevertheless, this study offers valuable insights into potential future directions by outlining the advantages and possible enhancements of our proposed method. One promising direction is the enhancement of PAtt-Lite's robustness, particularly concerning low-resource expressions, which could be achieved through further refinement of the patch extraction block. Additionally, to comprehensively evaluate the performance and robustness of our method, testing on additional FER datasets such as RAF-DB, FERPlus, and CK+ could provide valuable insights. Moreover, exploring and incorporating other methods into the architecture, such as zoning-based FER (ZFER), could potentially enhance performance without increasing complexity or computational costs.

Given EmoPAtt-Lite's lightweight nature, it holds promise for a wide range of applications deployable on edge or mobile devices. Beyond human-machine interaction scenarios, the model's utility extends to domains like healthcare and education, highlighting its versatility and potential for practical implementation.

REFERENCES

[1] A. Mehrabian, "Some referents and measures of nonverbal behavior," *Behavior Research Methods & Instrumentation*, vol. 1, pp. 203–207, 1968. [Online]. Available: https://api.semanticscholar.org/CorpusID:144744966

[2] Y. Wu and L. Shen, "An adaptive landmark-based attention network for students' facial expression recognition," in *2021 6th International Conference on Communication, Image and Signal Processing (CCISP)*, 2021, pp. 139–144.

[3] X. Li, R. Yue, W. Jia, H. Wang, and Y. Zheng, "Recognizing students' emotions based on facial expression analysis," in *2021 11th International Conference on Information Technology in Medicine and Education (ITME)*, 2021, pp. 96–100.

[4] C. Jonitta Meryl, K. Dharshini, D. Sujitha Juliet, J. Akila Rosy, and S. S. Jacob, "Deep learning based facial expression recognition for psychological health analysis," in *2020 International Conference on Communication and Signal Processing (ICCSP)*, 2020, pp. 1155–1158.

[5] G. Fu, Y. Yu, J. Ye, Y. Zheng, W. Li, N. Cui, and Q. Wang, "A method for diagnosing depression: Facial expression mimicry is evaluated by facial expression recognition," *Journal of Affective Disorders*, vol. 323, pp. 809–818, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016503272201388X

[6] J. L. Ngwe, K. M. Lim, C. P. Lee, and T. S. Ong, "Patt-lite: Lightweight patch and attention mobilenet for challenging facial expression recognition," 2023.

[7] Y. El Boudouri and A. Bohi, "Emonext: an adapted convnext for facial emotion recognition," in *2023 IEEE 25th International Workshop on Multimedia Signal Processing (MMSP)*, 2023, pp. 1–6.

[8] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, Y. Zhou, C. Ramaiah, F. Feng, R. Li, X. Wang, D. Athanasakis, J. Shawe-Taylor, M. Milakov, J. Park, R. Ionescu, M. Popescu, C. Grozea, J. Bergstra, J. Xie, L. Romaszko, B. Xu, Z. Chuang, and Y. Bengio, "Challenges in representation learning: A report on three machine learning contests," 2013.

[9] A. Adnan, M. Nawaz, S. Anwar, T. Ali, and M. Ali, "Object identification with color, texture, and object-correlation in cbir system," *World Academy of Science, Engineering and Technology*, vol. 64, pp. 117–122, 01 2010.

[10] S. Xie and H. Hu, "Facial expression recognition using hierarchical features with deep comprehensive multipatches aggregation convolutional neural networks," *IEEE Transactions on Multimedia*, vol. 21, no. 1, pp. 211–220, 2019.

[11] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, "Region attention networks for pose and occlusion robust facial expression recognition," 2019.

[12] S. Zhao, H. Cai, H. Liu, J. Zhang, and S. Chen, "Feature selection mechanism in cnns for facial expression recognition." in *BMVC*, vol. 12, 2018, p. 317.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015.

[14] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," 2014.

[15] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," 2017.

[16] 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.

[17] M. Aouayeb, W. Hamidouche, C. Soladie, K. Kpalma, and R. Seguier, "Learning vision transformer with squeeze and excitation for facial expression recognition," 2021.

[18] F. Xue, Q. Wang, and G. Guo, "Transfer: Learning relation-aware facial expression representations with transformers," 2021.

[19] F. Xue, Q. Wang, Z. Tan, Z. Ma, and G. Guo, "Vision transformer with attentive pooling for robust facial expression recognition," *IEEE Transactions on Affective Computing*, vol. 14, no. 4, p. 3244–3256, Oct. 2023. [Online]. Available: http://dx.doi.org/10.1109/TAFFC.2022.3226473

[20] S. Vignesh, M. Savithadevi, M. Sridevi, and R. Sridhar, "A novel facial emotion recognition model using segmentation vgg-19 architecture," *International Journal of Information Technology*, vol. 15, no. 4, pp. 1777–1787, 2023.

[21] N. Ali, A. Tubaishat, F. Al-Obeidat, M. Shabaz, M. Waqas, Z. Halim, I. Rida, and S. Anwar, "Towards enhanced identification of emotion from resource-constrained language through a novel multilingual bert approach," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, Apr. 2023, just Accepted. [Online]. Available: https://doi.org/10.1145/3592794

[22] A. U. Rahman, F. Al-Obeidat, A. Tubaishat, B. Shah, S. Anwar, and Z. Halim, "Discovering the correlation between phishing susceptibility causing data biases and big five personality traits using c-gan," *IEEE Transactions on Computational Social Systems*, vol. 11, no. 4, pp. 4800–4808, 2024.

[23] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," 2016.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023.