Sparse Attention to Emotion: Efficient Facial Emotion Recognition via Token Reduction

1st Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

2nd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

3rd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

***Abstract*—Facial Emotion Recognition (FER) is an important task that has significant implications across various fields such as biometrics, health, and human-computer interaction. Current Vision Transformer-based approaches display quadratic complexity $\mathcal{O}(N^2)$, with N being the input sequence length, making them cumbersome to deploy at the edge. In this paper, we hypothesize that the FER task does not necessarily require all facial information to correctly interpret emotional states, as specific regions such as the eyes, the mouth, and parts of the cheeks carry discriminative information that can be sufficient to recognize emotions. Based on this, we propose Sparse Attention to Emotion (SAE), a model that discards image tokens that have no added value to the emotional context, while preserving good accuracy and achieving a significant gain in computational cost. Surprisingly, even after suppressing 90% of the image tokens, our model achieves competitive accuracy to state of the art methods at much lower cost, providing a lightweight Facial Emotion Recognition approach. Extensive experimental results demonstrate that SAE achieves new state of the art results on the RAF-DB dataset while reducing the computational complexity by up to 90%.**

***Index Terms*—Vision Transformers, Model Compression, Token Pruning, Facial Emotion Recognition**

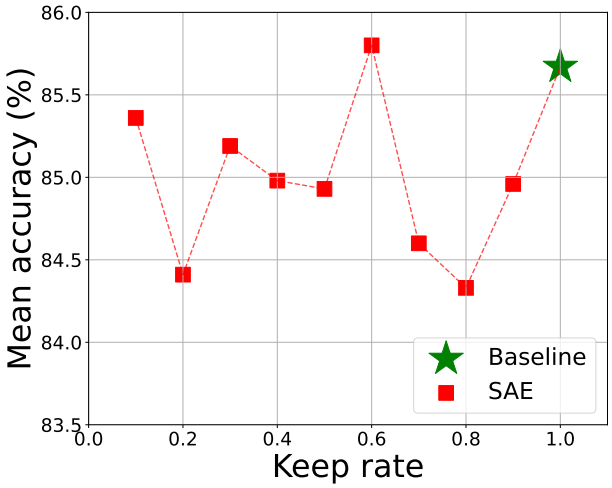

Fig. 1. Accuracy vs. keep rate of SAE on Raf-DB dataset.

I. Introduction

Facial Emotion Recognition (FER) has received great attention in recent years in various domains, including mental health [1], human-computer interaction [2], and virtual reality [3]. Translating human expressiveness into emotions is a fundamental step in several applications, from personalized experiences to psychological analysis [4], highlighting the importance of the research conducted.

Early FER approaches relied on handcrafted features for facial expression analysis. Although effective in controlled settings, such methods often lacked generalization and robustness in real world scenarios [5]. With the advent of deep learning, Convolutional Neural Networks (CNNs) were introduced to improve FER performance by automatically learning discriminative representations [6]–[8]. However, CNN based methods remain limited in capturing long range dependencies due to their inherently local receptive fields [9], [10].

To explicitly model the global contextual information of facial expressions, Vision Transformers (ViTs) [11], which have shown state of the art performance in image recognition, were the inspiration of the proposed Transformer-based FER models [12]–[14] by processing facial images as sequences of tokens. Among them, POSTER [13] achieves state of the art performance by integrating facial landmarks and image features through a two-stream pyramid cross-fusion architecture.

Despite these advances, the large number of image tokens treated indiscriminately to classify emotions, resulting in high computational costs, is a major limitation that remains unaddressed. In this work, we address this problem while maintaining high accuracy at a lower computational cost compared to state of the art methods [12]–[14].

We propose a token pruning method, originally used for generic image classification, that retains only the tokens with the highest impact on emotion recognition for the FER task and evaluates its performance on standard FER datasets. After experimenting with token pruning at different rates and analyzing the remaining tokens. We observe that in most cases the retained tokens correspond to specific regions of the face. Specifically, when retaining only 10% of the image tokens, which cover mostly the eyes and the mouth areas in later layers, our method still achieved state of the art

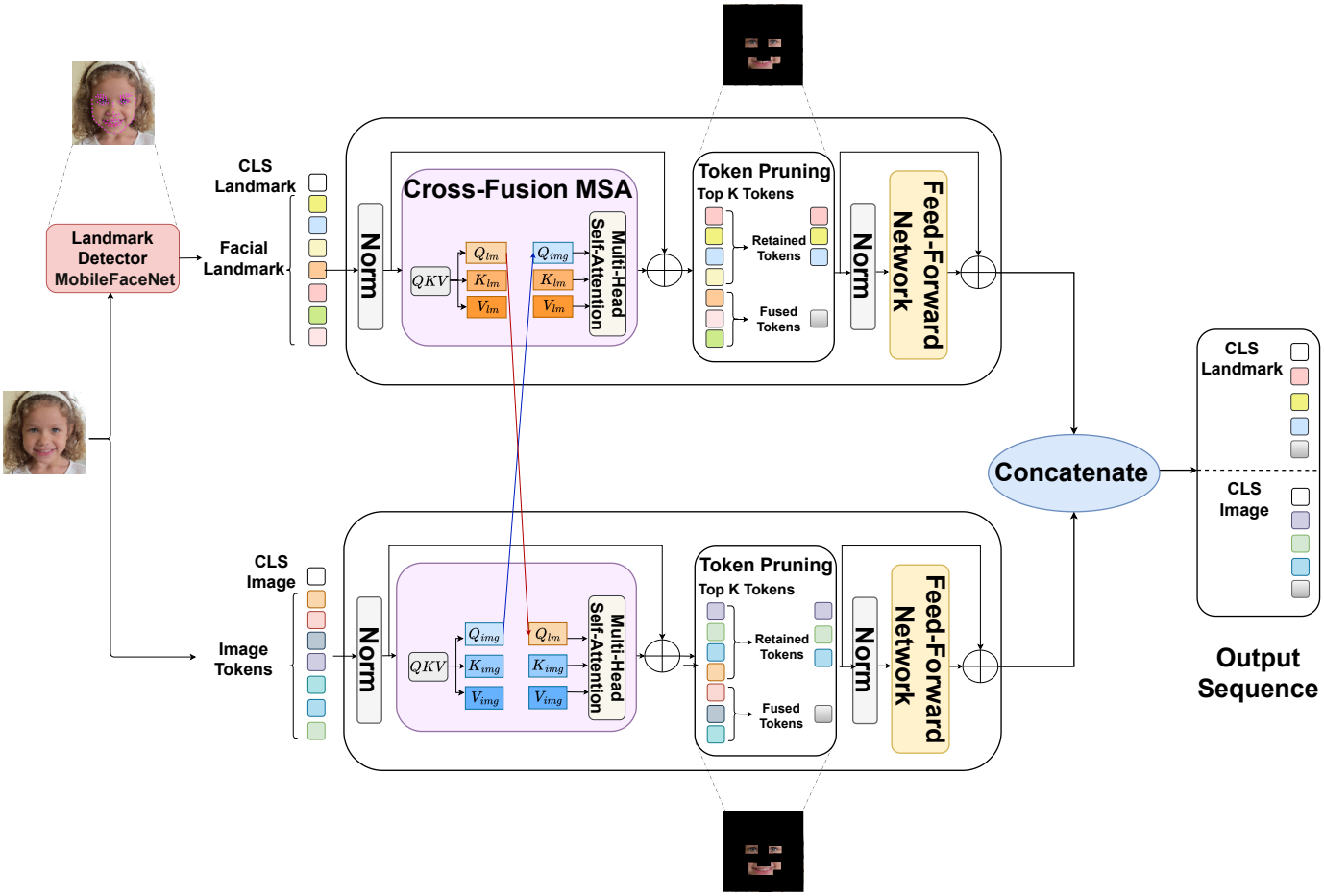

Fig. 2. Overview of SAE architecture.

performance. As shown in Figure 1, the best performance of SAE is achieved when keeping 60% of the image tokens, and even when discarding up to 90% of the tokens, SAE results in an accuracy drop of only 0.3% compared to the baseline, demonstrating its effectiveness even with aggressive token pruning. Therefore, our main finding is that the FER task does not necessarily need complete facial information, and that specific regions of the face, such as the eyes, mouth, and parts of the cheeks, are sufficient to predict emotional states. Based on this observation, we propose Sparse Attention to Emotion (SAE), a token pruning approach for the FER task. To the best of our knowledge, this is the first work to explore the application of token pruning as an efficient solution to Facial Emotion Recognition.

Empirical results on the Raf-DB [15] dataset show that we generally outperform state of the art methods on computational complexity reduction, while matching, or outperforming them in terms of accuracy. Specifically, our model achieves the best accuracy at a 20% token discard rate and a competitive performance even after discarding 90% of tokens. Our contributions are presented as follows:

- We propose a token pruning approach for the FER task to reduce the computational cost of ViT models, while limiting their performance drop.
- We simplify the complexity of FER by focusing on discriminative facial regions that carry important emotional information.
- We extensively validate the robustness and efficiency of our approach, demonstrating that it matches or outperforms previous state of the art methods on the RAF-DB dataset.

The remainder of the paper is organized as follows. In Section II, we review related work on facial emotion recognition and ViT-based approaches. Section III presents our proposed token pruning method for FER. Section IV details the experimental setup, datasets, and evaluation metrics. Finally, section V concludes the paper and discusses future directions.

II. Related works

A. Vision Transformers

The Vision Transformer (ViT) [11] represents an image as a sequence of patches and uses the attention mechanism to capture the correlations between them. DeiT [16] enhances data efficiency by introducing the knowledge distillation token, reducing the need for large-scale datasets required by ViT. Swin Transformer [17] introduces hierarchical representation

with a shifted window, limiting the attention computation to only local regions, while progressively merging patches. Pyramid Vision Transformer (PVT) [18] explores the multi-scale representation to capture features at different scales. CrossViT [19] leverages multi-scale feature fusion, enabling token interaction across different scales for better classification. LocalViT [20] reintegrates the convolutional local inductive bias into Transformers, improving local feature modeling without sacrificing global context. For a resource-constrained environment, MobileViT [21] introduces a mobile-friendly Vision Transformer by combining lightweight convolutions with Transformer layers, allowing efficient deployment.
Despite their effectiveness, these methods display quadratic complexity, as mentioned above, making them costly, which led to the development of compression techniques.

B. Token reduction

Multiple works addressed the complexity of Transformers through the reduction of the number of tokens processed. DynamicVit [22] introduces a module that prunes tokens across layers based on their importance. Building on this idea, EViT [23] proposes a simple and effective approach that removes inattentive tokens based on the cls-token attention score, significantly reducing the inference cost and without introducing additional parameters. Evo-ViT [24] introduces a slow-fast token evaluation strategy, in which informative tokens are selected and processed through all layers, while placeholder tokens evolve through fewer layers. ATS [25] presents a parameter-free module that adaptively selects informative tokens for each input image, leading to efficient processing by reducing computational cost while maintaining accuracy. SPViT [26] designs a latency-aware soft token pruning framework that selects the most informative tokens and integrates the less informative ones into a package token instead of discarding them. A-ViT [27] introduces an adaptive halting module that discards tokens once they reach their halting condition. Although many methods focus only on the token importance. The authors in [28] propose a method that considers both token importance and diversity.
Although these methods effectively reduce computational cost, their effectiveness is limited mainly to generic benchmarks such as ImageNet, and their applicability to domain-specific tasks such as Facial Emotion Recognition (FER) remains largely unexplored.

C. Facial Emotion Recognition (FER)

Facial Emotion Recognition (FER) is a task that identifies and classifies human emotions from facial expressions. Many approaches have been proposed to address this task, such as TransFER [12] which introduces a framework that learns relation-aware facial representations by modeling global interactions between local facial patches, while encouraging the diversity of learned facial features through attention dropping strategies (MAD, MSAD). VTFF [29] presents Visual Transformers with Feature Fusion, the method introduces an additional module ASF to combine multi-branch CNN features, and then applies a Transformer to capture global relationships among the resulting tokens, enhancing the robustness to occlusion and pose variation. EfficientFace [7] proposes a lightweight FER network by designing a local feature extractor and a channel spatial modulator. The model introduces a label distribution strategy to handle compound emotions, achieving robust performance under occlusion and pose variation. Face2Exp [30] presents a framework with a base network that learns general facial expressions and an adaptation network. This leads to a reduction in data bias and strong performance with less data. EAC [31] proposes a method for noisy-label FER by randomly erasing input images and using the flip attention consistency, preventing the model from overfitting to wrong labels. APViT [32] introduces two attentive pooling modules, APP is used to select the most attentive patches in CNN features, while APT is used to discard inattentive tokens in ViT, resulting in lower cost and better performance. DAN [33] designs facial emotion recognition frameworks consisting of three sub-networks that extract the features, attend to multiple facial regions, fuse features from different heads, and produce the prediction score. POSTER [13] introduces a FER framework that combines image features and facial landmark features, using two streams, and by doing a cross fusion attention, it uses a pyramid structure enabling multi-scale handling, addressing the intra-class discrepancy, inter-class similarity, and scale sensitivity. POSTER++ [5] proposes an improved version of POSTER, with the aim of reducing the expensive computational cost. ARBEx [34] introduces a ViT-based attentive feature extraction framework with a reliability balancing mechanism based on anchors, cross-attention, and confidence estimation. LFNSB [35] designs a facial recognition network that effectively balances model, complexity, and recognition accuracy, and this by using two key modules, LFN and SN. S2D [36] presents a framework that transfers knowledge from static facial expression recognition to dynamic facial expression. BTN [14] designs a framework in FER that learns reliable features by using class batch attention and multi-level attention to reduce noise and overfitting.

Compared to previous work that processes the entire image, our hypothesis is that not all facial regions are important for the FER task. Only a subset of facial regions is needed for the model to correctly classify an emotion, which is done by applying token pruning to remove unnecessary tokens and retain only the tokens of the most discriminative regions such as the eyes, the mouth, and parts of the cheeks. Our method not only lets the model focus more on the important and relevant facial regions but also reduces the computational cost.

III. Methodology

A. Self-Attention Mechanism

Given an input sequence $\mathbf{X} \in \mathbb{R}^{N \times d}$, where N is the number of tokens and d is the embedding dimension, the

TABLE I
COMPARISON OF SAE WITH STATE OF THE ART METHODS ON RAF-DB DATASET IN TERMS OF CLASS-WISE ACCURACY, AND MEAN ACCURACY AND KEEP RATE (TOKENS)

Method	Accuracy of Emotions (%)								Mean Acc (%)	keep rate
	Neutral	Happy	Sad	Surprise	Fear	Disgust	Anger	Contempt		
VTFF [29]	87.50	94.09	87.24	85.41	64.86	68.12	85.80	-	81.20	-
TransFER [12]	90.15	95.95	88.70	89.06	68.92	79.37	88.89	-	85.86	-
POSTER++ [5]	92.06	97.22	92.89	90.58	68.92	71.88	88.27	-	85.97	-
POSTER [13]	92.35	96.96	91.21	90.27	67.57	75.00	88.89	-	86.04	-
APViT [32]	92.06	97.30	88.70	93.31	72.97	73.75	86.42	-	86.36	-
BTN [14]	92.21	97.05	92.26	91.49	72.97	76.25	88.89	-	87.30	-
SAE	90.00	95.78	88.91	89.06	68.92	75.00	87.04	-	84.96	0.9
SAE	92.21	96.12	90.79	89.97	63.51	72.50	85.19	-	85.80	0.6
SAE	90.44	96.03	92.47	87.54	70.27	75.00	84.57	-	85.19	0.3
SAE	91.18	96.29	89.96	89.67	70.27	71.88	88.27	-	85.36	0.1

self-attention mechanism projects the input into queries, keys, values using learnable matrices $\mathbf{W_Q}, \mathbf{W_K}, \mathbf{W_V} \in \mathbb{R}^{d\times d}$

$$Q = XW_Q, \quad K = XW_K, \quad V = XW_V \tag{1}$$

The attention output is computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\Big(\frac{QK^\top}{\sqrt{d}}\Big)V. \tag{2}$$

The output of the self attention is then passed to a feed forward network (MLP):

$$\text{MLP}(X) = \text{FC}_2(\text{RELU}(\text{FC}_1(X))), \tag{3}$$

The total complexity cost of the Transformer layer is given by [37]:

$$\phi_{\text{BLK}}(N, d) = \phi_{\text{MSA}}(N, d) + \phi_{\text{MLP}}(N, d) = 12Nd^2 + 2N^2d, \tag{4}$$

highlighting the quadratic complexity with respect to the sequence length N.

B. Token Pruning

Given an input sequence $\mathbf{X} \in \mathbb{R}^{N\times d}$, where N is the number of tokens and d is the embedding dimension. We compute the self-attention matrix following Eq 2
We focus on the first row corresponding to the `[CLS]` token, each element a_i in this row representing the importance of the i-th token:

$$\mathbf{A}_{\text{CLS}} = \text{softmax}\Big(\frac{\mathbf{Q}_{\text{CLS}}\mathbf{K}^\top}{\sqrt{d}}\Big)V \tag{5}$$

$$a_i = (\mathbf{A}_{\text{CLS}})_i, \quad i = 1, \ldots, N. \tag{6}$$

For multi-head attention with H heads, the importance scores are averaged across all heads:

$$\bar{a} = \frac{1}{H}\sum_{h=1}^{H} a^{(h)}. \tag{7}$$

Let $\mathcal{K}$ denote the indices of the top-k tokens that are kept based on their importance scores, and $\mathcal{N}$ denote the indices of the remaining tokens. Inattentive tokens are fused into a single token, and the final token sequence is obtained by concatenating the kept tokens with the fused token. Formally:

$$\mathcal{K} = \text{TopK}(a_1, \ldots, a_N; k), \quad \mathcal{N} = \{1, \ldots, N\} \setminus \mathcal{K} \tag{8}$$

$$\mathbf{x}_{\text{fused}} = \sum_{i\in\mathcal{N}} a_i \mathbf{x}_i \tag{9}$$

$$\mathbf{X}_{\text{final}} = [\{\mathbf{x}_i\}_{i\in\mathcal{K}}, \mathbf{x}_{\text{fused}}] \tag{10}$$

C. Sparse Attention to Emotion

The overview of SAE is shown in Figure 2. Given an input facial image, we extract visual features using a convolutional backbone, while facial landmark features are obtained through a parallel landmark encoder. Let $\mathbf{X} \in \mathbb{R}^{N\times d}$, $\mathbf{L} \in \mathbb{R}^{N\times d}$ denote the image-based and landmark-based token embeddings, respectively, where N is the number of tokens and d is the embedding dimension.

the image and landmark sequences are processed by a cross-fusion multi-head self-attention (MSA) layer. Linear projections are first applied to obtain query, key, and value matrices as follows:

$$\mathbf{Q}_X = \mathbf{X}\mathbf{W}_Q^{(1)}, \quad \mathbf{K}_X = \mathbf{X}\mathbf{W}_K^{(1)}, \quad \mathbf{V}_X = \mathbf{X}\mathbf{W}_V^{(1)}, \tag{11}$$

$$\mathbf{Q}_L = \mathbf{L}\mathbf{W}_Q^{(2)}, \quad \mathbf{K}_L = \mathbf{L}\mathbf{W}_K^{(2)}, \quad \mathbf{V}_L = \mathbf{L}\mathbf{W}_V^{(2)}. \tag{12}$$

Cross fusion attention is then computed following Eq 2:

$$\text{ATTN}_X = \text{Attention}(\mathbf{Q}_L, \mathbf{K}_X, \mathbf{V}_X), \tag{13}$$

$$\text{ATTN}_L = \text{Attention}(\mathbf{Q}_X, \mathbf{K}_L, \mathbf{V}_L). \tag{14}$$

This cross-attention mechanism enables mutual information exchange between image and landmark representations.
Residual connections are applied to obtain the fused representations:

$$\mathbf{X}' = \text{ATTN}_X + \mathbf{X}, \tag{15}$$

$$\mathbf{L}' = \text{ATTN}_L + \mathbf{L}. \tag{16}$$

The fused sequences are then passed through the token pruning module:

$$\mathbf{X}^{pr} = \mathcal{TP}(\mathbf{X}'), \qquad \mathbf{L}^{pr} = \mathcal{TP}(\mathbf{L}'). \tag{17}$$

where $\mathcal{TP}$ represents the token pruning module. Finally, each pruned sequence is processed by a feed-forward network:

$$\mathbf{X}_{\text{out}} = \text{MLP}(\text{Norm}(\mathbf{X}^{pr})) + \mathbf{X}^{pr}, \tag{18}$$

$$\mathbf{L}_{\text{out}} = \text{MLP}(\text{Norm}(\mathbf{L}^{pr})) + \mathbf{L}^{pr}. \tag{19}$$

IV. Experiments

A. Datasets and experimental settings

RAF-DB: The Real-world Affective Faces Database (RAF-DB) [15] is a large-scale facial expression dataset containing real-world facial images. It includes seven emotion categories: surprise, fear, disgust, anger, neutral, sad, and happiness.

Training settings: The model is trained using the Sharpness Aware Minimization (SAM) optimizer wrapped around the Adam optimizer. The initial learning rate is set to 4×10^{-5}, with a batch size of 100, the learning rate is exponentially decayed using ExponentialLR with $\gamma = 0.98$ through the training process for 300 epochs. MobileFaceNet is used as the landmark feature extractor with frozen weights.

B. Experiment results

Table I presents the class-wise and mean accuracy of SAE on RAF-DB. For a detailed analysis of SAE performance compared to the state of the art, with a keep rate of 0.9, SAE achieves a mean accuracy of 84.96%, remaining close to state of the art methods that use full representations. When reducing the keep rate to 0.6, the mean accuracy increases to 85.60%, confirming that pruning can preserve discriminative tokens across most emotion classes. For keep rate of 0.3, despite discarding 70% of tokens, SAE maintains a competitive mean accuracy of 85.19%, with stable performance on dominant classes such as Happy, Neutral, and Sad. Even for a very low keep rate 0.1, where 90% of tokens are removed, the mean accuracy remains at 85.36%, a small degradation compared to the other approaches. Across all keep rates, our model shows strong robustness to token reduction, achieving significant low cost while preserving balanced per-class performance.

TABLE II
Comparison of SAE with state of the art approaches on RAF-DB in terms of overall accuracy (%) and keep rate (tokens).

Methods	Keep rate	overall accuracy
TransFER [12]	-	90.91
EfficientFace [7]	-	88.36
EAC [31]	-	90.35
APViT [32]	-	91.98
DAN [33]	-	89.70
POSTER [13]	-	92.05
POSTER++ [5]	-	92.21
ARBEx [34]	-	92.47
LFNSB [35]	-	91.07
S2D [36]	-	92.21
S2D* [36]	-	92.57
BTN [14]	-	92.54
SAE	0.9	90.51
SAE	0.6	91.17
SAE	0.3	91.00
SAE	0.1	91.13

Table II reports the overall accuracy of SAE compared to state of the art methods on RAF-DB, while focusing on the trade-off between performance and computational complexity. For a high keep rate 0.9, SAE reaches an accuracy of 90.51%, which is already comparable to other approaches. When the keep rate is reduced to 0.6, the accuracy increases to 91.13%, remaining competitive with recent transformer-based methods. Even with more aggressive pruning for a keep rate of 0.3 and 0.1, SAE shows strong robustness, the accuracy remains stable. These results clearly indicate that SAE can reduce computational complexity by up to 90% while causing only a negligible performance loss often below 1% demonstrating an efficient accuracy complexity trade-off compared to existing approaches.

V. Conclusion

In this paper, we proposed Sparse Attention to Emotion (SAE), an efficient approach that explores token pruning in Facial Emotion Recognition (FER) to reduce computational complexity without compromising performance. SAE integrates a pyramid cross-fusion architecture that exploits complementary information from landmark and image streams, and uses multi-scale feature representation to capture facial emotions at different resolutions. Extensively, experiments on common benchmark datasets showed that SAE matches or outperforms state of the art results while retaining only 10% of the image tokens, making it well suited for edge deployment and resource constrained applications.

References

[1] N. Shanthi, A. A. Stonier, A. Sherine, T. Devaraju, S. Abinash, R. Ajay, V. Arul Prasath, and V. Ganji, "An integrated approach for mental health assessment using emotion analysis and scales," *Healthcare Technology Letters*, vol. 12, no. 1, p. e12040, 2025.

[2] U. Chindiyababy, P. Kakkar, J. Vedula, J. Yunus, A. Umidbek, and S. Sharma, "Deep learning-based facial emotion recognition for advanced human-computer interaction," in *2025 First International Conference on Advances in Computer Science, Electrical, Electronics, and Communication Technologies (CE2CT)*. IEEE, 2025, pp. 1247–1252.

[3] Q. Gong, X. Liu, and Y. Ma, "Real-time facial expression recognition based on image processing in virtual reality," *International Journal of Computational Intelligence Systems*, vol. 18, no. 1, p. 8, 2025.

[4] D. Ruan, Y. Yan, S. Lai, Z. Chai, C. Shen, and H. Wang, "Feature decomposition and reconstruction learning for effective facial expression recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 7660–7669.

[5] J. Mao, R. Xu, X. Yin, Y. Chang, B. Nie, and A. Huang, "Poster++: A simpler and stronger facial expression recognition network," 2023. [Online]. Available: https://arxiv.org/abs/2301.12149

[6] M. Akhand, S. Roy, N. Siddique, M. A. S. Kamal, and T. Shimamura, "Facial emotion recognition using transfer learning in the deep cnn," *Electronics*, vol. 10, no. 9, p. 1036, 2021.

[7] Z. Zhao, Q. Liu, and F. Zhou, "Robust lightweight facial expression recognition network with label distribution training," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 4, 2021, pp. 3510–3519.

[8] A. V. Savchenko, "Facial expression and attributes recognition based on multi-task learning of lightweight neural networks," in *2021 IEEE 19th international symposium on intelligent systems and informatics (SISY)*. IEEE, 2021, pp. 119–124.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[10] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

[11] 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 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[12] F. Xue, Q. Wang, and G. Guo, "Transfer: Learning relation-aware facial expression representations with transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 3601–3610.

[13] C. Zheng, M. Mendieta, and C. Chen, "Poster: A pyramid cross-fusion transformer network for facial expression recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2023, pp. 3146–3155.

[14] M. Her, J. Jeong, H. Song, and J.-H. Han, "Batch transformer: Look for attention in batch," *IEEE Access*, vol. 13, p. 190093–190107, 2025. [Online]. Available: http://dx.doi.org/10.1109/ACCESS.2025.3628323

[15] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2852–2861.

[16] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International conference on machine learning*. PMLR, 2021, pp. 10 347–10 357.

[17] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[18] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 568–578.

[19] C.-F. R. Chen, Q. Fan, and R. Panda, "Crossvit: Cross-attention multi-scale vision transformer for image classification," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 357–366.

[20] Y. Li, K. Zhang, J. Cao, R. Timofte, M. Magno, L. Benini, and L. Van Goo, "Localvit: Analyzing locality in vision transformers," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 9598–9605.

[21] S. Mehta and M. Rastegari, "Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer. arxiv 2021," *arXiv preprint arXiv:2110.02178*, 2021.

[22] Y. Rao, W. Zhao, B. Liu, J. Lu, J. Zhou, and C.-J. Hsieh, "Dynamicvit: Efficient vision transformers with dynamic token sparsification," *Advances in neural information processing systems*, vol. 34, pp. 13 937–13 949, 2021.

[23] Y. Liang, C. GE, Z. Tong, Y. Song, J. Wang, and P. Xie, "EVit: Expediting vision transformers via token reorganizations," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=BjyvwnXXVn_

[24] Y. Xu, Z. Zhang, M. Zhang, K. Sheng, K. Li, W. Dong, L. Zhang, C. Xu, and X. Sun, "Evo-vit: Slow-fast token evolution for dynamic vision transformer," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, no. 3, 2022, pp. 2964–2972.

[25] M. Fayyaz, S. A. Koohpayegani, F. R. Jafari, S. Sengupta, H. R. V. Joze, E. Sommerlade, H. Pirsiavash, and J. Gall, "Adaptive token sampling for efficient vision transformers," in *European conference on computer vision*. Springer, 2022, pp. 396–414.

[26] Z. Kong, P. Dong, X. Ma, X. Meng, W. Niu, M. Sun, X. Shen, G. Yuan, B. Ren, H. Tang *et al.*, "Spvit: Enabling faster vision transformers via latency-aware soft token pruning," in *European conference on computer vision*. Springer, 2022, pp. 620–640.

[27] H. Yin, A. Vahdat, J. M. Alvarez, A. Mallya, J. Kautz, and P. Molchanov, "A-vit: Adaptive tokens for efficient vision transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 809–10 818.

[28] S. Long, Z. Zhao, J. Pi, S. Wang, and J. Wang, "Beyond attentive tokens: Incorporating token importance and diversity for efficient vision transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10 334–10 343.

[29] F. Ma, B. Sun, and S. Li, "Facial expression recognition with visual transformers and attentional selective fusion," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1236–1248, 2021.

[30] D. Zeng, Z. Lin, X. Yan, Y. Liu, F. Wang, and B. Tang, "Face2exp: Combating data biases for facial expression recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 20 291–20 300.

[31] Y. Zhang, C. Wang, X. Ling, and W. Deng, "Learn from all: Erasing attention consistency for noisy label facial expression recognition," in *European Conference on Computer Vision*. Springer, 2022, pp. 418–434.

[32] 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, pp. 3244–3256, 2022.

[33] Z. Wen, W. Lin, T. Wang, and G. Xu, "Distract your attention: Multi-head cross attention network for facial expression recognition," *Biomimetics*, vol. 8, no. 2, p. 199, 2023.

[34] A. T. Wasi, K. Šerbetar, R. Islam, T. H. Rafi, and D.-K. Chae, "Arbex: Attentive feature extraction with reliability balancing for robust facial expression learning," *arXiv preprint arXiv:2305.01486*, 2023.

[35] X. Chen and L. Huang, "A lightweight model enhancing facial expression recognition with spatial bias and cosine-harmony loss," *Computation*, vol. 12, no. 10, p. 201, 2024.

[36] Y. Chen, J. Li, S. Shan, M. Wang, and R. Hong, "From static to dynamic: Adapting landmark-aware image models for facial expression recognition in videos," *IEEE Transactions on Affective Computing*, 2024.

[37] Z. Pan, B. Zhuang, J. Liu, H. He, and J. Cai, "Scalable vision transformers with hierarchical pooling," in *Proceedings of the IEEE/cvf international conference on computer vision*, 2021, pp. 377–386.