Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Compact Vision Transformer with Pyramid and Landmarks Fusion for Facial Expression Recognition

Yuecheng Yin

yinyuecheng@peu.edu.cn

Puer University

Xiuzhong Tang

Puer University

Gen Ye

Puer University

Jiaxu Dao

Puer University

Research Article

Keywords: facial expression recognition, deep learning, vision Transformer, facial landmarks, emotion classification

Posted Date: August 14th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-7302219/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A Compact Vision Transformer with Pyramid and Landmarks Fusion for Facial Expression Recognition

Yuecheng Yin*
Pu'er Universityr
No. 6 Xueyuan Road
Pu'er, China
yinyuecheng@peu.edu.cn

Xiuzhong Tang
Pu'er Universityr
No. 6 Xueyuan Road
Pu'er, China
tangxiuzhong@peu.edu.cn

Gen Ye
Pu'er Universityr
No. 6 Xueyuan Road
Pu'er, China
yegen@peu.edu.cn

Jiaxu Dao
Pu'er Universityr
No. 6 Xueyuan Road
Pu'er, China
daojiaxu@peu.edu.cn

ABSTRACT
This paper proposes the FacialViT, a lightweight and efficient facial expression recognition model based on EfficientViT. It combines multi-scale image features and facial landmark information to enhance recognition accuracy while maintaining low computational cost. A Feature Pyramid Fusion Module improves robustness to expression scale variation, and a Cross-Attention Fusion Block enables interaction between global and local features. FacialViT achieves 74.9% accuracy on FER2013 with only 0.53 GFLOPs, outperforming many existing methods. Its efficiency and effectiveness make it well-suited for real-time applications on edge devices and mobile platforms.

General Terms
Pattern Recognition;Algorithms;Computer Vision

Keywords
facial expression recognition, deep learning, vision Transformer, facial landmarks, emotion classification.

1. INTRODUCTION
Facial expression recognition (FER), as an important technology in the field of artificial intelligence, has received widespread attention in recent years. It aims to improve the intelligent level of human-computer interaction by analyzing the emotional information contained in static images or dynamic video sequences [1]. Facial expression is the main way for humans to express emotions, and its non-verbal characteristics make it play a vital role in interpersonal communication. Compared with traditional communication methods such as sound, gestures and body language, facial expressions can convey richer emotional information and transcend cultural and language barriers. Therefore, recognizing and analyzing facial expressions is of great significance for understanding human emotions, improving human-computer interaction and promoting the development of artificial intelligence[2], [3]. Facial expression recognition (FER) technology can be widely used in the security governance of digital society. It can analyze the emotions of people at large-scale events, help government departments improve the level of social governance, maintain public safety, and provide safer and more convenient services to the public[4] .

In the early stages of FER technology development, traditional computer vision algorithms based on handcrafted features were mainly used to achieve this[5], [6]. For example, commonly used feature extraction algorithms include Histogram of Oriented Gradients (HOG)[7], SIFT (Scale-Invariant Feature Transform)[8], etc. However, these traditional computer vision algorithms have high computational complexity and the algorithms themselves lack sufficient robustness and accuracy, so the overall effect of FER is also poor.

In recent years, with the rapid advancement of technologies such as deep learning , FER research has also made significant progress. Researchers have developed various models based on deep learning. Convolution neural networks (CNN)[9], graph convolution networks (GCN)[10] and Transformer[11] have all been applied to FER and have proposed corresponding solutions to different challenges. In addition, new methods such as multi-modal fusion, self-supervised learning and weakly supervised learning have also injected new vitality into the development of FER technology. These algorithms focus on changes in facial appearance, such as wrinkles and folds, and can extract features from the entire face or specific areas , and then classify these features . The performance of the algorithm mainly depends on the process of extracting features. The training and

inference processes of these models require a lot of computing resources, and the deployment cost is expensive. The algorithm itself is sensitive to the scale of the face and is difficult to be truly applied to actual scenarios with low latency requirements. Recently, some work has also tried to solve the problems of lightweight and low latency of FER models[12],[13] , which has certain reference significance for the research of this paper.

This paper introduces a lightweight Transformer-based FER network, FacialViT. The model adopts the lightweight EfficientViT[14] architecture, integrates facial landmarks, and uses cross-fusion of multi-scale features, taking into account both recognition accuracy and model computational efficiency, making the model more suitable for deployment in embedded devices and other high-performance application scenarios, and promoting the application of FER technology on a large number of mobile devices. The main work can be summarized as follows:

(1)The paper proposed a lightweight model FacialViT based on Transformer , which uses multi-scale face pyramid feature fusion to solve the scale sensitivity problem of faces in the recognition process.

(2)The EfficientViT architecture is used as the backbone network to extract image features, achieving high-quality feature extraction while ensuring the lightweight of the model . The extracted features of different scales are fused through the Pyramid Feature Fusion Module .

(3)The face landmark detection was introduced, and the Cross-Attention Fusion Block was proposed to cross-fuse face image features and landmark features. This can better capture the detailed changes in facial expressions and improve the accuracy and robustness of facial expression recognition.

(4)The results verified the efficiency and accuracy of the proposed FacialViT. The FacialViT-Mx series model achieved an accuracy of 74.9% on FER, surpassing most SOTA methods, while the computational workload was only 0.53GFLOPs, far less than these SOTA algorithms, achieving a good balance between computational efficiency and recognition accuracy.

2. RELATED RESEARCH

In this section, we briefly review some recent studies on facial emotion classification . With the rapid development of deep learning in the field of computer vision, these techniques have been found to be increasingly useful for the challenging facial expression recognition (FER) task[16], [17]. Wang et al.[18] proposed a regional attention network (RAN) to capture facial regions with occlusion and pose variations for FER task.

Georgescu et al.[19] proposed a method in which automatic features learned by multiple CNN architectures are combined with handcrafted features computed using a bag-of-visual-words (BOVW) model . After the fusion of these two types of features , the category label of each test image can be predicted using a local learning framework. Pecoraro et al.[20] introduced a new channel self-attention module called Local Multi-Head Channel (LHC) that can be seamlessly integrated into any existing CNN architecture. This module overcomes the limitations of the global attention mechanism by focusing on key facial features that have a significant impact on facial expressions. In another study [21], Fard et al. proposed an adaptive correlation (Ad-Corre) loss consisting of three components: feature discriminator (FD), mean discriminator (MD), and embedding discriminator (ED). The proposed Ad-Corre loss is combined with the traditional cross-entropy loss to train two backbone models: Xception[22] and Resnet50[23]. As the Transformer network has made great breakthroughs in the field of natural language processing (NLP) , it has also had a great impact on the field of computer vision . In NLP, the Transformer is designed to model long sequence inputs. When applied to computer vision tasks, ViT[24] divides the image into image patches, and then the self-attention mechanism in the Transformer can capture the long-range dependencies between these image patches. EfficientViT is an efficient vision model based on Vision Transformer (ViT), which aims to solve the bottleneck problems of traditional ViT models in terms of computational complexity and memory usage. In the FER task, Aouayeb et al.[25] directly applied the ViT structure and added a SE module before the MLP layer to perform the FER task. Xue et al.[26] proposed a Transformer-based method named TransFER , which designed a local CNN module to locate different local image patches and a Transformer encoder to explore the global relationship between these local image patches.

Current research primarily focuses on improving recognition accuracy, often leading to large model sizes and high computational complexity. These models impose considerable computational burdens in real-time applications, making them unsuitable for large-scale deployment on embedded devices.

Compared with previous work, FacialViT employs a pyramid-based feature fusion strategy to address the challenge of scale sensitivity in facial expression recognition (FER) tasks. By leveraging the EfficientViT architecture, the model enhances both the lightweight design and computational efficiency, achieving a better trade-off between recognition accuracy and computational complexity. This design facilitates deployment in real-time scenarios, especially those with constrained computational resources.

3. PROPOSED METHODS

3.1 Overview of FacialViT Architecture

The FacialViT architecture comprises three key components:e

(1) an EfficientViT module for extracting local attention features;p

(2) a Feature Pyramid Fusion Module (FPFM) for multi-scale feature fusion, which enhances recognition accuracy across varying facial expression scales;a

(3) a Cross-Attention Fusion Block (CAFB) that integrates global image features with local facial landmark features, and strengthens their interaction via a self-attention mechanism.p

These components collectively improve the accuracy and robustness of facial expression recognition. The overall architecture is illustrated in Figure 1.

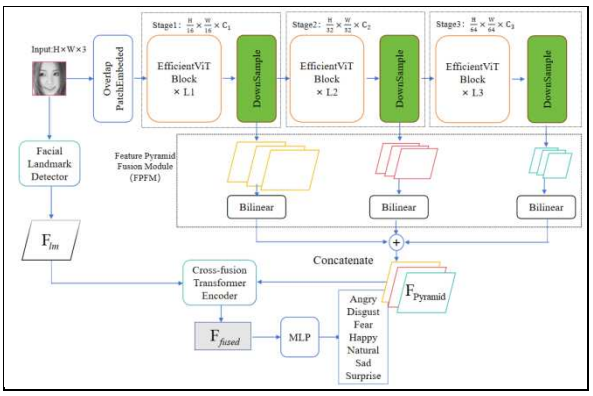


Fig1: Overview of the FacialViT Architecture

3.2 Preprocessing

In the preprocessing stage, MTCNN[27] was employed for face detection and alignment. Subsequently, data augmentation was applied to improve model generalization. The augmentation techniques included random horizontal flipping, random rotation, random affine transformation, and Gaussian blur. These methods were primarily used to mitigate overfitting caused by the limited size of the dataset. Finally, all face images were normalized.

3.3 Backbone

In this study, EfficientViT is adopted as the backbone network for feature extraction in facial expression recognition. EfficientViT is a lightweight model that integrates the Vision Transformer (ViT) architecture with efficient computational strategies. It significantly reduces computational and storage overhead while maintaining high accuracy. Compared with traditional convolutional neural networks (CNNs), EfficientViT offers a better trade-off between computational efficiency and representational capacity, making it particularly suitable for deployment in resource-constrained environments.

The core advantage of EfficientViT lies in its use of sparsification and attention mechanisms, which effectively capture key information from images while reducing redundant computations. In facial expression recognition tasks, the backbone network must accurately and efficiently extract subtle differences in facial expressions—an area where EfficientViT excels. Its efficient feature extraction capability enables the model to quickly adapt to expression variations, enhancing both recognition accuracy and real-time performance.

Moreover, the lightweight nature of EfficientViT allows deployment on hardware-constrained devices such as embedded systems and mobile terminals, further extending its applicability and practical value.

3.4 Feature Pyramid Fusion Module

In facial expression recognition tasks, facial features at different scales contain rich and complementary information. Low-level features from shallower layers capture fine-grained textures, such as wrinkles and subtle changes around the eyes, while high-level features from deeper layers encode more abstract semantic information, such as overall facial deformation. Relying on single-scale features may result in information loss.

The Feature Pyramid Network (FPN), originally proposed by Lin et al.[6], addresses this issue by extracting and fusing features at multiple scales to enhance the recognition of small targets. Inspired by this approach, we design a Feature Pyramid Fusion Module (FPFM) tailored for facial expression recognition. The FPFM effectively integrates multi-scale facial features to improve the robustness and accuracy of expression recognition. The architecture of the proposed FPFM is illustrated in Figure 2.

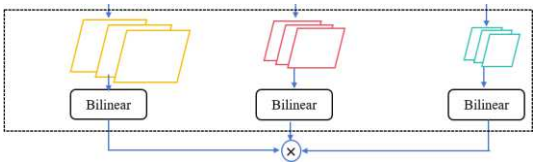


Fig2: Feature Pyramid Fusion Module (FPFM)

In the FPFM structure, we assume that the backbone network outputs L feature maps at different scales, denoted as:

$$F = \{F_1, F_2, \ldots, F_L\}, \quad F_i \in \mathbb{R}^{C_i \times H_i \times W_i} \tag{1}$$

Where C_i is the number of channels in the i-th layer, H_i, W_i is the spatial size of the feature map. To align multi-scale features, bilinear interpolation is applied to resize all feature maps to a uniform spatial resolution, denoted as:

$$\tilde{F}_i = \text{Upsample}(F_i, H_T, W_T), \quad \forall i \in \{1, 2, \ldots, L\} \tag{2}$$

H_T, W_T is the size of the intermediate layer feature, thes features are concatenated in the channel dimension (Channel-wise):

$$F_{\text{FPF}} = \text{Concat}(\tilde{F}_1, \tilde{F}_2, \ldots, \tilde{F}_L), \quad F_{\text{FPF}} \in \square^{(C_1+C_2+\ldots+C_L)\times H_T \times W_T} \tag{3}$$

The final fused features F_{FPF} were fed into the FER classifier. Compared with single-scale expression recognition methods, the proposed Feature Pyramid Fusion Module (FPFM) fully leverages multi-scale information, enhancing the model's robustness to variations in face scale, viewing angles, and lighting conditions.

3.5 Cross-Attention Fusion Block

In the facial expression recognition (FER) task, it is usually difficult to effectively capture the subtle changes of facial expressions by simply extracting the features of the face image , especially in the face of diverse facial movements and complex lighting conditions. To overcome this limitation, we use MobileFaceNet[28] to detect facial landmarks and fuse the landmark information with the face image information to improve the reliability of recognition. Inspired by POSTER[29], this study proposes a new feature cross-attention fusion module that combines the global features of the image with the local key point features, and strengthens the interactive fusion between the two through the self-attention mechanism, thereby improving the accuracy and robustness of facial expression recognition.

Cross-Attention Fusion Block aims to utilize the mutual guidance and information fusion between the image feature stream and the key point feature stream , and cross-fuse the features of the two in multiple dimensions through the self-attention mechanism. The main idea of this module is that the global information of the image provides contextual information for the key points, while the local features of the key points help the image features focus on the key areas of expression recognition, such as eyes and mouth. In this way, the model can dynamically adjust the attention to different areas through cross-attention and improve the sensitivity to expression changes. As shown in Figure 3, F_{lm} is the key point information feature and F_{Pyramid} is the fused image pyramid feature .

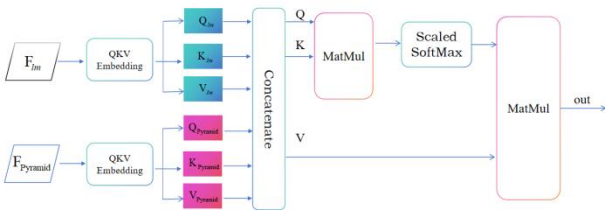


Fig3: Our proposed Cross-Attention Fusion Block

By performing linear transformations on image features and key point features, query , key , and value matrices are generated. The image feature matrix $F_{\text{Pyramid}} \in \mathrm{R}^{N\times D}$, $F_{lm} \in \mathrm{R}^{N\times D}$, where N represents the number of patches of the feature, and D represents the dimension of the feature. Image features and key point features are transformed through the weight matrix W_Q , W_K , W_V to generate query, key, and value, respectively:

$$Q_{\text{Pyramid}} = F_{\text{Pyramid}} W_Q, K_{\text{Pyramid}} = F_{\text{Pyramid}} W_K, V_{\text{Pyramid}} = F_{\text{Pyramid}} W_V \tag{4}$$

$$Q_{lm} = F_{lm} W_Q, K_{lm} = F_{lm} W_K, V_{lm} = F_{lm} W_V \tag{5}$$

After concatenate operation, Q, K, V were generated as follows:

$$\begin{cases} Q = Concate(Q_{\text{Pyramid}}, Q_{lm}) \\ K = Concate(K_{\text{Pyramid}}, K_{lm}) \\ V = Concate(V_{\text{Pyramid}}, V_{lm}) \end{cases} \tag{6}$$

Calculated through the cross-attention mechanism, the output information is as follows:

$$out = SoftMax(\frac{QK^T}{\sqrt{d}})V \tag{7}$$

The final output features are passed through an MLP classifier to complete the facial expression recognition process. The Cross-Attention Fusion Block integrates multi-scale global image information with local key point features, enabling the model to better capture expression changes, especially subtle facial movements. The attention mechanism guides the model to dynamically adjust its focus on different areas of expression changes, thereby improving sensitivity to key expression regions. This module effectively handles challenges such as lighting variations, occlusions, and posture changes, enhancing the model's stability in dynamic environments.

4. EXPERIMENTS

4.1 FER Dataset

This study conducts experiments on the FER2013 dataset, which was introduced in the 2013 ICML Representation Learning Challenge[30]. The dataset contains 35,887 grayscale face images, each with a resolution of 48×48 pixels (see Figure 4). It is divided into three subsets: 28,709 images for training, 3,589 images for validation, and 3,589 images for testing. The images are labeled into seven basic emotion categories: Angry, Disgust, Fear, Happy, Neutral, Sad, and Surprise. With its large-scale sample size, FER2013 is widely used in facial expression recognition (FER) tasks. However, the dataset exhibits a significant class imbalance problem, which presents a challenge for training deep learning models. To mitigate this issue, the research introduces a data augmentation strategy during the preprocessing stage. Example images from the FER2013 dataset are shown in Figure 4.



Fig 4: Sample images from the FER2013 dataset

4.2 Training

This paper implemented FacialViT in an end-to-end manner using PyTorch[31] on an NVIDIA GTX1070 GPU. we adopted EfficientViT14 as the image backbone network, which was pre-trained on the ImageNet-1K dataset[32]. The backbone weights were updated during the training process of FacialViT. For the facial landmark detection, we chose MobileFaceNet28, and its parameters remained frozen during training. In the Pyramid Feature Fusion Module, we extracted multi-scale features from the backbone, with feature maps at resolutions (H/16, W/16), (H/32, W/32), and (H/64, W/64) corresponding to the large, medium, and small scales, respectively, which were then fused accordingly. For the Cross-Attention Fusion Block, the Q (Query), K (Key), and V (Value) matrices of the facial landmark features and global image features were fused via element-wise concatenation. we set the batch size to 48 and the learning rate to 1e-4, with a cosine decay schedule to enhance convergence. Additionally, this research incorporated data augmentation techniques, such as random cropping and random rotation, to augment the training data and improve the model's ability to generalize.

4.3 Results

The results presented in Table 1 demonstrate that our proposed model, FacialViT, outperforms many existing FER network architectures on the FER2013 dataset. Among the listed models, FacialViT-M5 achieves the best performance, with an accuracy of 74.9%.

This achievement is primarily attributed to the unique design and architecture of FacialViT, which effectively integrates facial image features with key point features to better capture subtle facial expression changes. Additionally, the multi-scale feature fusion design enhances the model's robustness across different facial scales. In terms of accuracy, FacialViT-M5 achieves a state-of-the-art level, outperforming well-known models such as ResNet50 (73.20%) and VGG (73.28%), and reaching performance comparable to cutting-edge models like EmoNeXt35 and Mini-ResEmoteNet44. Notably, FacialViT-M5 maintains a low computational complexity of only 0.53 GFLOPs, making it significantly lighter than other models with similar performance.

We also propose a more lightweight variant, FacialViT-M0, which incurs a computational cost of only 0.09 GFLOPs and achieves real-time performance on many devices while maintaining a competitive accuracy of 69.3%. These results demonstrate that the proposed FacialViT model effectively captures and classifies facial emotions with excellent accuracy–efficiency trade-offs, making it well-suited for deployment in resource-constrained environments.

Table 1. Accuracy and FLOPs comparison of the FER-2013 test set

Model	Accuracy	GFLOPs
Segmentation VGG-19[33]	75.97	19.63
Local Learning Deep+BOW[34]	75.42	×
EmoNeXt[35]	76.12	4.55
VGG[36]	73.28	15.47
CNN Hyperparameter Optimisation[37]	72.16	2.8
Ad-Corre[38]	72.03	×
Mini-ResEmoteNet(B)[39]	70.20	×
DeepEmotion[40]	70.02	~ 0.16
Local Learning BOW[41]	67.48	×
ResEmoteNet[42]	79.79	4.33
ResidualMaskingNetwork[43]	74.14	21.93
Mini-ResEmoteNet(A)[44]	76.33	×
GoogleNet[45]	65.20	1.50
Inception-V3[46]	71.60	5.73
ResNet50[47]	73.20	4.11
FacialViT-M0	69.3	0.09
FacialViT-M5	74.9	0.53

This paper proposed the FacialViT-M series of models, which achieve an improved trade-off between model accuracy and computational complexity (measured in GFLOPs). Figure 5 presents a comparative analysis between the FacialViT series and mainstream FER models, including the VGG and Inception series, in terms of GFLOPs and Top-1 Accuracy. As illustrated in the figure, FacialViT-M0 to FacialViT-M5 demonstrate a progressive improvement in accuracy while maintaining extremely low computational overhead, highlighting their strong scalability and efficiency.

In particular, FacialViT-M5 achieves 74.9% accuracy with only 0.53 GFLOPs, significantly outperforming traditional models such as InceptionV3 and VGG-16 at similar or even higher computational costs. Despite its lightweight design, the FacialViT series delivers performance comparable to, or even exceeding, that of high-FLOPs models, confirming its potential for deployment in resource-constrained environments,

especially in edge computing and mobile applications where lightweight models are essential.

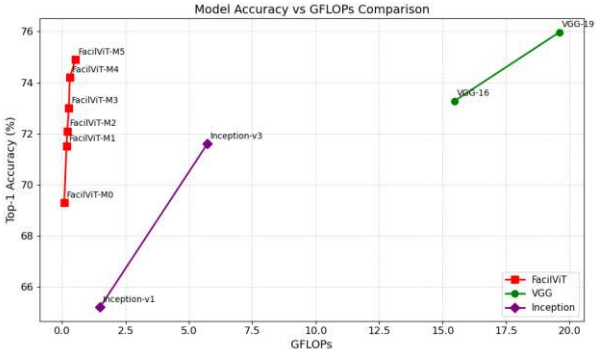


Fig5: Model Accuracy VS Computation Load

5. CONCLUSIONS

This paper proposes FacialViT, a lightweight deep learning framework for facial expression recognition (FER) based on the EfficientViT architecture. The proposed framework incorporates a Cross-Attention Fusion Block, which enables deep integration of facial landmark information with visual features, thereby enriching facial representations and enhancing the accuracy of emotion classification. In addition, a Feature Pyramid Fusion Module (FPFM) is designed to effectively integrate multi-scale features extracted by EfficientViT, further improving the model's ability to recognize expressions across varying spatial resolutions.

Experimental results on the FER2013 dataset demonstrate that FacialViT outperforms many current mainstream deep learning models in terms of classification accuracy. By seamlessly integrating the Feature Pyramid Fusion Module (FPFM), Cross-Attention Fusion Block, and the EfficientViT backbone, FacialViT achieves a highly lightweight design while maintaining high recognition accuracy. The model exhibits strong and efficient performance in facial expression recognition tasks, highlighting its broad potential for real-world applications, particularly in resource-constrained environments.

Furthermore, the research is going to extend the application of FacialViT to practical scenarios such as digital social governance, further exploring its real-world value and expanding its potential research directions.

6. ACKNOWLEDGMENTS

This work was supported by Academic Research Projects of Pu'er University PEXYXJYB202320.

7. FUNDING

This research was supported by Academic Research Projects of Pu'er University under Grant Number PEXYXJYB202320.

8. REFERENCES

[1] O. Rudovic, J. Lee, M. Dai, B. Schuller, and R. W. Picard, "Personalized machine learning for robot perception of affect and engagement in autism therapy," Science Robotics, vol. 3, no. 19, p. eaao6760, 2018.

[2] J. Z. Wang, S. Zhao, C. Wu, R. B. Adams, M. G. Newman, T. Shafir, and R. Tsachor, "Unlocking the emotional world of visual media: An overview of the science, research, and impact of understanding emotion drawing insights from psychology, engineering, and the arts." Proceedings of the IEEE,vol. 111, no. 10, pp. 1236–1286, 2023.

[3] Y. Wang, W. Song, W. Tao, A. Liotta, D. Yang, X. Li, S. Gao, Y. Sun, W. Ge, W. Zhang, and W. Zhang, "A systematic review on affective computing: Emotion models, databases, and recent advances," Information Fusion, vol. 83, pp. 19–52, 2022.

[4] Wu J .Cyberspace Endogenous Security and Safety Defense Vision[M].Springer, Singapore,2024.

[5] Guoying Zhao and Matti Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE transactions on pattern analysis and machine intelligence, 2007. 1.

[6] Lin Zhong, Qingshan Liu, Peng Yang, Bo Liu, Junzhou Huang, and Dimitris N Metaxas. Learning active facial patches for expression analysis. In 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012.

[7] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). Ieee, 2005.

[8] David G Lowe. Distinctive image features from scale invariant keypoints. International journal of computer vision, 60(2):91–110, 2004.

[9] J. Cai, Z. Meng, A. S. Khan, Z. Li, J. O'Reilly, and Y. Tong, "Island loss for learning discriminative features in facial expression recognition," in 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition(FG 2018). IEEE, 2018, pp. 302–309.

[10] T. Liu, J. Li, J. Wu, B. Du, J. Chang, and Y. Liu, "Facial expression recognition on the high aggregation subgraphs," IEEE TIP, vol. 32, pp. 3732–3745, 2023.

[11] C. Zheng, M. Mendieta, and C. Chen, "Poster: A pyramid cross-fusion transformer network for facial expression recognition," in ICCVW, 2023, pp. 3138–3147.

[12] M. Jeong, J. Nam, and B. C. Ko, “Lightweight multilayer random forests for monitoring driver emotional status,” Ieee Access, vol. 8, pp. 60 344–60 354, 2020.

[13] Y. Shang, M. Yang, J. Cui, L. Cui, Z. Huang, and X. Li, “Driver emotion and fatigue state detection based on time series fusion,” Electronics, vol. 12, no. 1, p. 26, 2023.

[14] X. Liu, H. Peng, N. Zheng, Y. Yang, H. Hu, and Y. Yuan, “Efficientvit: Memory efficient vision transformer with cascaded group attention,” in Proceedings of the IEEE/CVF Conference on Computer Vision andPattern Recognition, 2023, pp. 14 420–14 430.

[15] Lin T Y , Dollar P , Girshick R ,et al.Feature Pyramid Networks for Object Detection[J].IEEE Computer Society, 2017.DOI:10.1109/CVPR.2017.106.

[16] Yuhang Zhang, Chengrui Wang, and Weihong Deng. Relative uncertainty learning for facial expression recognition. Advances in Neural Information Processing Systems, 34:17616–17627, 2021. 2

[17] Yuhang Zhang, Chengrui Wang, Xu Ling, and WeihongDeng. Learn from all: Erasing attention consistency for noisy label facial expression recognition. In Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI, pages 418–434. Springer, 2022. 2.

[18] Kai Wang, Xiaojiang Peng, Jianfei Yang, Debin Meng, and Yu Qiao. Region attention networks for pose and occlusion robust facial expression recognition. IEEE Transactions on Image Processing, 29:4057–4069, 2020. 1, 2, 6

[19] M.-I. Georgescu, R. T. Ionescu, and M. Popescu, “Local learning with deep and handcrafted features for facial expression recognition,” IEEE Access, vol. 7, pp. 64 827–64 836, 2019.

[20] R. Pecoraro, V. Basile, and V. Bono, “Local multi-head channel selfattention for facial expression recognition,” Information, vol. 13, no. 9, p. 419, 2022

[21] A. P. Fard and M. H. Mahoor, “Ad-corre: Adaptive correlation-based loss for facial expression recognition in the wild,” IEEE Access, vol. 10, pp.26 756–26 768, 2022.

[22] F. Chollet, “Xception: Deep learning with depthwise separable convolutions,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1251–1258.

[23] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[24] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition atscale. ICLR, 2021. 3, 4

[25] Mouath Aouayeb, Wassim Hamidouche, Catherine Soladie, Kidiyo Kpalma, and Renaud Seguier. Learning vision transformer with squeeze and excitation for facial expression recognition. arXiv preprint arXiv:2107.03107, 2021. 3

[26] Fanglei Xue, Qiangchang Wang, and Guodong Guo. Transfer: Learning relation-aware facial expression representations with transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 3601–3610, October 2021. 2, 3, 5, 6, 9, 10

[27] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, “Joint face detection and alignment using multitask cascaded convolutional networks,” IEEE signal processing letters, vol. 23, no. 10, pp. 1499–1503, 2016

[28] Cunjian Chen. PyTorch Face Landmark: A fast and accurate facial landmark detector, 2021. 3, 4, 5

[29] Zheng C , Mendieta M , Chen C .POSTER: A Pyramid Cross-Fusion Transformer Network for Facial Expression Recognition[J]. 2022.DOI:10.48550/arXiv.2204.04083.

[30] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee et al., “Challenges in representation learning: A report on three machine learningcontests,” in Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3-7, 2013. Proceedings,Part III 20. Springer, 2013, pp. 117–124.

[31] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 5

[32] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In CVPR, 2009. 1, 5, 6, 7, 8

[33] Vignesh S, Savithadevi M, Sridevi M, et al. A novel facial emotion recognition model using segmentation VGG-19 architecture[J]. International Journal of Information Technology, 2023, 15(4): 1777-1787.

[34] Georgescu M I, Ionescu R T, Popescu M. Local learning with deep and handcrafted features for facial expression recognition[J]. IEEE Access, 2019, 7: 64827-64836.

[35] El Boudouri Y, Bohi A. Emonext: an adapted convnext for facial emotion recognition[C]//2023 IEEE 25th International Workshop on Multimedia Signal Processing (MMSP). IEEE, 2023: 1-6.

[36] Khaireddin Y, Chen Z. Facial emotion recognition: State of the art performance on FER2013[J]. arXiv preprint arXiv:2105.03588, 2021.

[37] Vulpe-Grigoraşi A, Grigore O. Convolutional neural network hyperparameters optimization for facial emotion recognition[C]//2021 12th International Symposium on Advanced Topics in Electrical Engineering (ATEE). IEEE, 2021: 1-5.

[38] Fard A P, Mahoor M H. Ad-corre: Adaptive correlation-based loss for facial expression recognition in the wild[J]. IEEE Access, 2022, 10: 26756-26768.

[39] Murtada A, Abdelrhman O, Attia T A. Mini-ResEmoteNet: Leveraging Knowledge Distillation for Human-Centered Design[J]. arXiv preprint arXiv:2501.18538, 2025.

[40] Minaee S, Minaei M, Abdolrashidi A. Deep-emotion: Facial expression recognition using attentional convolutional network[J]. Sensors, 2021, 21(9): 3046.

[41] Goodfellow I J, Erhan D, Carrier P L, et al. Challenges in representation learning: A report on three machine learning contests[C]//Neural information processing: 20th international conference, ICONIP 2013, daegu, korea, november 3-7, 2013. Proceedings, Part III 20. Springer berlin heidelberg, 2013: 117-124.

[42] Roy A K, Kathania H K, Sharma A, et al. ResEmoteNet: bridging accuracy and loss reduction in facial emotion recognition[J]. IEEE Signal Processing Letters, 2024.

[43] Pham L, Vu T H, Tran T A. Facial expression recognition using residual masking network[C]//2020 25Th international conference on pattern recognition (ICPR). IEEE, 2021: 4513-4519.

[44] Murtada A, Abdelrhman O, Attia T A. Mini-ResEmoteNet: Leveraging Knowledge Distillation for Human-Centered Design[J]. arXiv preprint arXiv:2501.18538, 2025.

[45] P. Giannopoulos, I. Perikos, and I. Hatzilygeroudis, “Deep learning approaches for facial emotion recognition: A case study on fer-2013,” Advances in Hybridization of Intelligent Methods: Models, Systems and Applications, pp. 1–16, 2018.

[46] Y. Khaireddin and Z. Chen, “Facial emotion recognition: State of the art performance on fer2013,” arXiv preprint arXiv:2105.03588, 2021.

[47] A. Khanzada, C. Bai, and F. T. Celepcikay, “Facial expression recognition with deep learning,” arXiv preprint arXiv:2004.11823, 2020.