Proceedings of the 4th International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/2026.33778

Application of Facial Expression Recognition in Real-World Environments: From Traditional Convolutional Networks to Vision Transformers

Zhenling Feng
College of Computer Science and Artificial Intelligence, Shanxi Normal University, Taiyuan, China
2351020239@sxnu.edu.cn

Abstract. Facial expression recognition (FER), as the core technology in the cross field of emotional computing and computer vision, has important theoretical significance and practical application value in the real environment. It not only promotes the theoretical deepening of emotional computing, human-computer interaction and other related disciplines, but also provides emotional perception support for multi domain scenes, and helps to realize a more intelligent and humanistic interaction mode. Firstly, this paper briefly describes the research significance and application value of facial expression recognition in real environment; Then, the evolution of FER technology is sorted out, and the technical iteration from traditional convolutional neural network (CNN) to vision transformer (VIT) is analyzed. The unique advantages of the attention mechanism supported by ViT in capturing global features and modeling long-distance dependence are clarified, as well as its specific mechanism in dealing with complex challenges such as light change, face occlusion, pose shift and so on in the real environment; Then, it systematically summarizes the core challenges faced by FER research in the current real environment, and sorts out the frontier research directions and development trends such as multimodal fusion; Finally, the key problems that have not been solved in this field are pointed out, and the future technology development trend is prospected.

Keywords: Facial expression recognition, Convolutional neural network, Vision Transformer, Attention mechanism, Multimodal fusion

1. Introduction

Facial expression recognition (FER), as the core technology in the cross field of emotional computing and computer vision, has irreplaceable application value in the fields of human-computer interaction, intelligent monitoring, medical diagnosis, driving safety warning and so on. It not only promotes the theoretical deepening of emotional computing, human-computer interaction and other related disciplines, but also provides accurate perceptual support for multi domain scenes and helps to achieve a more intelligent interaction mode. With the upgrading of industrial landing demand, FER research has shown the inevitable trend of transformation from laboratory environment to in the wild environment. The performance of high-precision models in laboratory environment has

DOI: 10.54254/2755-2721/2026.33778

significantly declined in complex real scenes. How to improve the robustness and generalization of models under non ideal conditions has become the core research proposition in the field. Systematically combing the evolution of FER technology and solving the bottleneck of real environment landing has important theoretical and practical significance for promoting FER technology [1]. In recent years, research in the field of FER has made leaps and bounds, and a number of authoritative works have been published. Wang et al. Systematically sorted out the technical system of static and dynamic FER, identified the core challenges such as expression interference, cross domain inconsistency, and composite emotions in real scenes, and built a complete analytical framework for field research [2]. For the technical pain points in the real scene, Liu et al. Proposed a dynamic expression recognition method based on facial motion unit enhancement, which effectively improved the recognition effect under unbalanced samples [3]; Aiming at the problem of fuzzy labels of real data, Zhang et al. Proposed an adaptive sample mining algorithm, which achieved excellent recognition accuracy on real scene data sets such as RAF-DB [4]. At the same time, ViT has become the core direction of FER technology iteration. Relevant research focuses on lightweight design and multi-scale feature fusion optimization. While enhancing the robustness of the model, it continuously reduces the computational cost to adapt to the actual deployment requirements [5].

At present, the real environment FER still faces many core technical problems, and its scene category covers complex non ideal conditions such as dramatic changes in lighting, posture angle offset, partial occlusion of face, low-resolution imaging, cross domain migration, etc. However, there are still a lot of scientific and engineering problems to be solved in the direction of efficient fusion of multimodal data, lightweight and landing adaptation of transformer model. Existing FER reviews focus on the evolution of the technology architecture itself, but generally ignore the logical relationship between technology progress and actual deployment, and fail to fully respond to the core needs of the industry. Based on this, this review takes real-world applications as the core orientation, systematically sorts out the complete development of FER technology from CNN to ViT, and deeply analyzes the advantages, disadvantages and adaptation boundaries of different technical routes in real scenes, aiming to bridge the gap between model innovation and application, and clearly solve the complex problems in the real world is the core driving force to promote the sustainable development of FER technology.

2. FER basic theory and research preparation

2.1. Basic framework of facial expression recognition

This study is based on Ekman's six basic emotional frameworks (happiness, anger, sadness, surprise, fear, disgust) in the physiological characteristics and emotional classification of facial expressions, as well as the classification of compound emotions based on the basic emotions. FER is divided into static FER (SFER) (based on image) and dynamic FER (DFER) (based on video). The core difference between the two is that DFER needs to solve the problems of key frame extraction, spatio-temporal feature capture, expression intensity temporal change and so on.

FER is systematically divided into two categories: image-based static FER (SFER) and video-based dynamic FER (DFER), and their challenges and solutions are sorted out. The research difference lies in the challenges faced by DFER, such as key frame extraction, spatio-temporal feature extraction, expression intensity changes and so on, which makes it essentially different from SFER in model design and methodology [2].

DOI: 10.54254/2755-2721/2026.33778

2.2. FER related technical basis

2.2.1. Core principles of convolutional neural networks

$$f(x,y) = (I^{*}K)(x,y) = \sum_{m}\sum_{n} I(x-m, y-n)K(m,n) \tag{1}$$

Where I is the input image and K is the convolution kernel. The local texture feature is captured through the sliding window, but there are some problems such as limited receptive field and insufficient global relationship modeling.

Convolutional neural network is composed of convolution layer, pooling layer, activation layer and full connection layer. It realizes hierarchical feature extraction by sharing local receptive fields and weights, captures edge texture in shallow layer, and learns semantic structure in shallow layer. The pooling operation is used to reduce the dimension and enhance the translation invariance of the model. Convolution and pooling stacking structure are often used in network design, and residual connection and multi-scale fusion are combined to improve the performance of feature expression and recognition [6].

2.2.2. Basic idea of attention mechanism

The mechanism aims to make the model automatically focus on key areas, suppress irrelevant information, and simulate the selective attention of human vision. In facial expression recognition, the model can automatically focus on the expression sensitive areas such as eyebrows and eyes, corners of the mouth, etc. by paying attention, so as to improve the perception ability of subtle expression changes. Its core is to get attention weight by calculating the similarity of query, key and value. Taking self attention as the core, calculate the similarity of query vector (Q), key vector (K) and value vector (V) to achieve feature weighting. The formula is:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^{T}}{\sqrt{d_k}}\right)V \tag{2}$$

In the transformer structure, multiple attention maps features into multiple sets of Q , K , V parallel computing and splicing to improve the ability of feature expression [7].

2.2.3. Core architecture of ViT

The ViT model first evenly divides the input face image into multiple image blocks of fixed size, and flattens them into one-dimensional sequence features [8]. Then, the block features are mapped into the model input vector through the embedding layer, and the learnable position embedding is superimposed to retain the spatial structure information of the image. The obtained sequence features are sent to the multi-layer transformer encoder, and then iterated feature extraction is carried out through the multi head self attention mechanism and feed-forward network [9]. Among them, the multi head self attention can model the long-range dependence between different regions of the face, and the feedforward network can further carry out the nonlinear transformation of features, and finally realize the joint modeling of the global structure of facial expression and key local features.

DOI: 10.54254/2755-2721/2026.33778

2.2.4. Basic methods of multimodal fusion

Multimodal fusion is an important enhancement method in facial expression recognition (FER). It can make up for the deficiency of single visual mode by combining multi-source information such as facial image, voice and text, and improve the recognition robustness in complex scenes. According to the fusion stage, it can be divided into three categories: data level fusion directly splices the original multimodal data, and the information is intact but vulnerable to noise; Feature level fusion stitches or adaptively weights the extracted expression features to achieve complementary enhancement between modes; Decision level fusion uses voting or weighted fusion results in the model output stage, which is more flexible and generalization.

3. Evolution of FER Technology: from traditional algorithm to convolutional network

3.1. Classic methods and evaluation of CNN

The traditional FER era belongs to the manual driven mode, and its core relies on researchers' manual design of feature descriptors based on psychology and prior knowledge. The generalization accuracy of the results obtained by such methods is limited. With the breakthrough of AlexNet, deep CNN began to popularize in the field of FER and officially entered the era of deep learning. Represented by classic networks such as VGG and ResNet, CNN uses convolution to obtain local end-to-end facial features, as shown in figure 1, The level by level abstraction and classification has achieved remarkable results in initial accuracy, so it has become the mainstream method for FER in laboratory environment [1]. However, only scanning facial expressions based on local convolution kernel will result in sensory field limitation, which can not effectively capture long-distance dependent pain points in facial regions such as eyebrows and eyes, corners of mouth, etc., and its performance will be significantly reduced in real environment scenes with occlusion and posture changes.

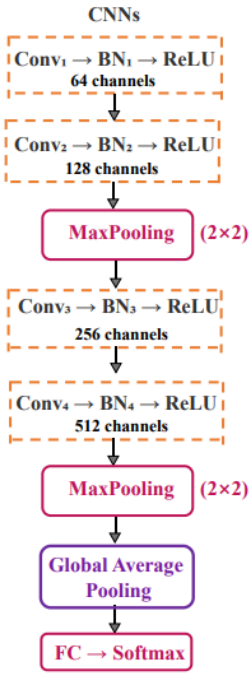

Figure 1. CNNs architecture (picture credit: original)

DOI: 10.54254/2755-2721/2026.33778

3.2. Rise of ViT

In order to be universal in the real non laboratory environment and solve the problem of global dependence, ViT began to rise. ViT converts two-dimensional images into one-dimensional sequences through image block embedding. In fact, it cuts the face into multiple image blocks. The multi head self attention mechanism can infer the occluded area by using the unobstructed area, and calculate the relationship between all blocks at the same time, so as to macro model the global features of the face, as shown in Figure 2, It can accurately capture the regional correlation of complex and subtle expressions, showing better robustness on SFER data sets in real environments (such as AffectNet), and solving the problems such as CNN's lack of global relationship modeling and sensitivity to local posture changes [10]. In the actual deployment and landing, it is found that the use of ViT architecture has large data requirements, reduced performance in small samples and low resolution, high computational complexity, and cross domain generalization ability to be improved.

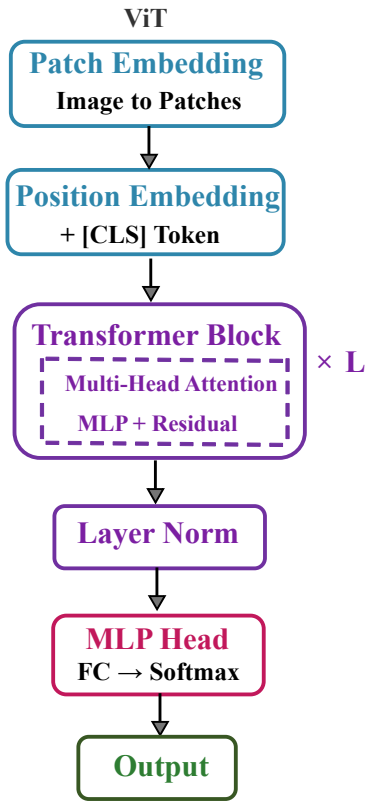

Figure 2. ViT architecture (picture credit: original)

3.3. CNN-ViT hybrid model

Combining CNN's local detail texture extraction ability with ViT's modeling ability of correlation and global dependency, ViT has optional local perception ability through soft convolution inductive bias, squeeze ViT uses double token fusion, adding compression module between layers to realize gradual dimension reduction, patch attention mechanism to enhance the robustness of occlusion and other designs, balancing the model accuracy and computational efficiency of pure CNN or pure ViT [11], which has become the mainstream technology direction of FER in the real environment. The typical design is the fusion of CNN's front-end local feature extraction and ViT's back-end global

DOI: 10.54254/2755-2721/2026.33778

feature, as shown in Fig. 3. therefore, it shows that FER is developing towards high efficiency and lightweight.

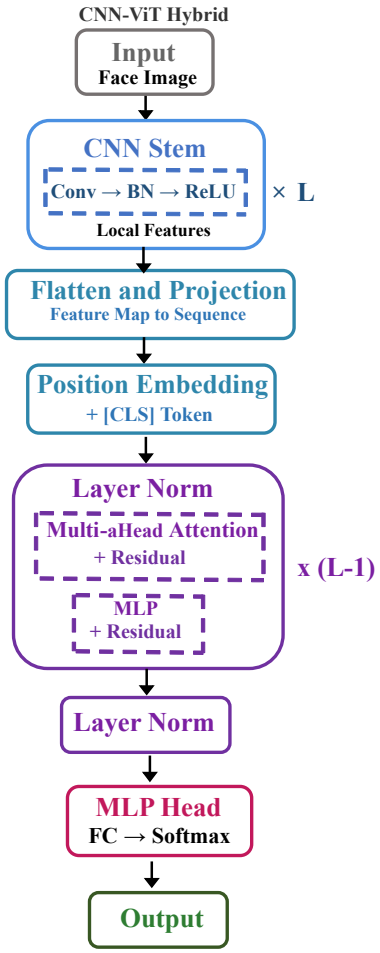

Figure 3. CNN-ViT Hybrid architecture (picture credit: original)

4. Resource support and ViT optimization application of real environment FER

4.1. FER data set and core evaluation system

The research conducted a comprehensive cross dataset analysis and generalization experiment on the mainstream facial expression data sets, covering comparative analysis

Laboratory data sets (CK+, JAFFE, MMI) are used for benchmark testing to verify the basic ability of the model, while real environment data sets (AffectNet, FER2013, RAF-DB) are used for robustness evaluation to verify the actual availability of the model. Taking accuracy rate (ACC), recall rate (REC), F1 value and cross domain generalization accuracy as the core evaluation indexes, the comprehensive evaluation of model performance is realized [12]. Among them, the accuracy rate is used to reflect the accuracy of the overall classification of the model, which is the most intuitive global indicator to evaluate the recognition effect; Recall rate focuses on the recognition completeness of the model for all kinds of expressions, especially those with few samples and easily confused expressions, so as to avoid evaluation bias caused by data imbalance; The comprehensive accuracy and recall rate of F1 value can reflect the robustness and classification reliability of the model in complex scenes in a more balanced way. On this basis, the cross domain generalization

DOI: 10.54254/2755-2721/2026.33778

accuracy is introduced to test the migration ability of the model in different data sets, different acquisition environments and different face distribution, so as to comprehensively verify the adaptability and stability of the model in real complex scenes. Through the joint evaluation of multi-dimensional indicators, the systematic analysis of model recognition accuracy, category balance and generalization ability can be realized to ensure that the evaluation results are more comprehensive and credible.

4.2. Adaptation and performance for dynamic and static FER

Based on the above core evaluation indicators, the specific adaptation method of ViT in FER task is discussed. The original architecture includes image blocking strategy, feature embedding optimization, transformer encoder, etc. The optimization block strategy of ViT model realizes the FER adaptation transformation. For the transformation of FER task, the performance advantage analysis on the dynamic and static data sets in the real environment.

4.2.1. Lightweight ViT design

To solve the problem of performance degradation under small samples/low resolution, it is necessary to balance the accuracy and efficiency of ViT, and solve the problem of high ViT computing cost. At present, the academia adopts lightweight ViT design optimization strategies such as model compression, shift window, and multi-layer knowledge distillation

Through the quantization of ViT model, the model parameters are compressed from high precision (such as 32-bit floating point number) to low precision (such as 8-bit integer), as well as pruning, knowledge distillation and other optimization technologies, which reduce the amount of calculation, solve the problem of high calculation cost, and facilitate deployment. Experiments also show that the optimized ViT can significantly reduce the computational overhead while maintaining the accuracy; Or the fixed weight classification layer, which does not update the weight during back propagation, significantly reduces the training time and solves the high cost of ViT training, which is also an innovative lightweight training scheme [13]; The swin transformer divides the image into non overlapping local windows and only calculates self attention in the window. The next layer moves the window by half to make the information flow between different windows, reducing the complexity from $\left(O\left(n^2\right)\right)$ to $(O(n))$[14]. As an effective model compression strategy, multi-layer knowledge distillation makes the lightweight student model fit the characteristic distribution of the teacher model in multiple middle layers, which not only reduces the amount of model parameters and computational overhead, but also maintains the original recognition performance.

4.2.2. CNN-ViT hybrid architecture for sheltered environment adaptation

Static real environment often has occlusion scenes (masks, sunglasses, etc.), so it is difficult to correctly identify the input image. It is necessary to combine the advantages of CNN and ViT, while taking into account local features and global dependence, and balancing accuracy and efficiency.

Because of the different performance of CNN and ViT hybrid models in FER, academia analyzed how to balance local feature extraction and global dependency modeling, and balance accuracy and efficiency. CoatNet provides a framework for the combination of convolution and attention. The soft convolution inductive bias mechanism is introduced to enhance the ability of the model to capture local features [15]. At present, the convolutional visual transformer designed for occluded scenes

DOI: 10.54254/2755-2721/2026.33778

enhances the robustness to occlusion through patch attention mechanism; ViT combination of slider locally weighted convolutional attention and global attention pooling.

4.2.3. ViT improvement for dynamic FER

The ViT model, which integrates frame level emotion guidance mechanism, can effectively capture the continuous temporal changes and dynamic evolution of facial expressions in video sequences while maintaining the fine modeling ability of spatial features. By applying independent emotional supervision and feature guidance to the expression features of each frame, this kind of method strengthens the perception ability of the model for subtle expression changes, and uses transformer's temporal attention to model the long-range dependence, so as to realize the joint modeling of dynamic expression from spatial structure to temporal evolution, and significantly improve the accuracy and robustness in natural scene video expression recognition.

5. Conclusion

This paper systematically reviews the evolution of facial expression recognition (FER) technology from manual drive to data drive, from CNN local modeling to hybrid architecture global dependency, reveals the core laws of feature extraction from manual design to end-to-end automatic learning, and the continuous expansion of modeling range from local texture to global dependency, and analyzes the advantages of CNN in the laboratory environment and the fundamental limitations of limited receptive field and insufficient long-distance dependency capture in the real environment. On this basis, it is explained that CNN-ViT hybrid architecture combines the local and global advantages, and lightweight design, dynamic timing expansion and other improvement directions promote the practical process of ViT in the real environment.

Future FER research will focus on the two aspects of multimodal fusion and cross domain generalization. Multimodal fusion needs to explore data types and fusion logic, sort out methods such as feature level and decision level, and overcome the problems of heterogeneity and modal missing; Cross domain generalization requires the development of domain adaptive technology to improve the generalization ability of the real scene of the model. The specific research paths include: using massive unlabeled data for pre training, improving the generalization ability of the model in small samples and cross domain scenarios, and realizing self supervised and weakly supervised learning; Create an efficient and well aligned multimodal fusion framework, explore the cross modal interaction mechanism, solve the problems of heterogeneity and temporal asynchrony, and support the deployment of real-time systems; In terms of privacy protection, it is oriented to personalized, fair and ethical deployment, and promotes the development of FER system to be transparent, credible and people-centered.

References

[1] Sarvakar, K., & Rana, K. (2025). Revolutionizing facial emotion recognition: In-depth analysis of cutting-edge models, methodologies, and datasets. Discover Artificial Intelligence, 5(1), 388. https: //doi.org/10.1007/s44163-025-00553-w

[2] Wang, Y., Yan, S., Liu, Y., Song, W., Liu, J., Chang, Y., Mai, X., Hu, X., Zhang, W., & Gan, Z. (2024). A Survey on Facial Expression Recognition of Static and Dynamic Emotions (arXiv: 2408.15777). arXiv. https: //doi.org/10.48550/arXiv.2408.15777

[3] Liu, F., Gu, L., Shi, C., & Fu, X. (2025). Action Unit Enhance Dynamic Facial Expression Recognition (arXiv: 2507.07678). arXiv. https: //doi.org/10.48550/arXiv.2507.07678

DOI: 10.54254/2755-2721/2026.33778

[4] Liu, X., An, L., Shen, K., Zhang, Z., & Sun, X. (2025). Fuzzy Membership-Driven Adaptive Sample Mining for Facial Expression Recognition. IEEE Transactions on Fuzzy Systems, 33(12), 4240–4251. https: //doi.org/10.1109/TFUZZ.2025.3593951.
[5] Yanqiu, L., Shengzhao, L., Guangling, S., & Pu, Y. (2025). Lightweight Swin Transformer combined with multi-scale feature fusion for face expression recognition. Opto-Electronic Engineering, 52(1), 240234–14. https: //doi.org/10.12086/oee.2025.240234
[6] Lu, H., & Zhang, Q. (2016). Application research review of deep convolutional neural networks in computer vision. Journal of Data Acquisition and Processing, 31(1), 1–17. https: //doi.org/10.16337/j.1004-9037.2016.01.001
[7] Tiberi, L., Mignacco, F., Irie, K., & Sompolinsky, H. (2024). Dissecting the Interplay of Attention Paths in a Statistical Mechanics Theory of Transformers (arXiv: 2405.15926). arXiv. https: //doi.org/10.48550/arXiv.2405.15926
[8] Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., Yang, Z., Zhang, Y., & Tao, D. (2023). A Survey on Vision Transformer. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(1), 87–110. https: //doi.org/10.1109/TPAMI.2022.3152247
[9] DosoViTskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (arXiv: 2010.11929). arXiv. https: //doi.org/10.48550/arXiv.2010.11929
[10] Chaudhari, A., Bhatt, C., Adiraju, A., & Mazzeo, P. L. (2022). ViTFER: Facial Emotion Recognition with Vision Transformers. Applied System Innovation, 5, 80. https: //doi.org/10.3390/asi5040080
[11] Kim, S., Nam, J., & Ko, B. C. (2022). Facial Expression Recognition Based on Squeeze Vision Transformer. Sensors, 22(10), 3729. https: //doi.org/10.3390/s22103729
[12] Li, S., & Deng, W. (2022). A Deeper Look at Facial Expression Dataset Bias. IEEE Transactions on Affective Computing, 13(2), 881–893. https: //doi.org/10.1109/TAFFC.2020.2973158
[13] Xu, Y., Duan, X., Fan, P., Zhao, Z., & Guo, X. (2025). Combining Fixed-Weight ArcFace Loss and Vision Transformer for Facial Expression Recognition. Sensors, 25(23), 7166. https: //doi.org/10.3390/s25237166
[14] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows (arXiv: 2103.14030). arXiv. https: //doi.org/10.48550/arXiv.2103.14030
[15] d'Ascoli, S., Touvron, H., LeaViTt, M., Morcos, A., Biroli, G., & Sagun, L. (2022). ConViT: Improving Vision Transformers with Soft Convolutional Inductive Biases. Journal of Statistical Mechanics: Theory and Experiment, 2022(11), 114005. https: //doi.org/10.1088/1742-5468/ac9830