An Investigation on Facial Expression Recognition Using Convolutional Neural Networks

Yitong Chen*

College of Information Science and Engineering, Hunan Normal University, 410081 Changsha, China

Abstract. Facial expressions serve as a fundamental non-verbal channel for conveying human emotions and intentions. This review provides a systematic examination of Convolutional Neural Network (CNN)-based methods for Facial Expression Recognition (FER). To address challenges inherent in real-world scenarios, including occlusions, pose variations, and the subtle nature of expressions, research focus has shifted from standard network architectures to more specialized designs. The paper elaborates on three primary technical paradigms: firstly, attention-based CNN architectures, such as channel and spatial attention mechanisms, which enhance robustness by dynamically focusing on critical facial regions; secondly, multi-modal fusion approaches that compensate for the limitations of relying solely on RGB data by integrating geometric features, thermal imaging, or depth information, proving particularly effective for handling occlusions; and finally, strategies for integrating local and global features, including multi-pathway networks and 3D CNNs, which effectively capture hierarchical information ranging from fine-grained muscle movements to holistic expressive dynamics. Despite continuous methodological innovations, the field continues to grapple with multiple challenges, including data bias, insufficient environmental robustness, difficulties in micro-expression recognition, and computational efficiency demands. Looking ahead, key directions for advancing FER towards more reliable, efficient, and practical deployment involve integrating causal reasoning with symbolic knowledge, developing dynamic adaptive inference frameworks.

1 Introduction

Facial expressions serve as one of the most fundamental and natural non-verbal channels for conveying human emotions and intentions. Automated FER technology aims to enable computers to understand and classify these emotional signals, demonstrating significant potential across various domains such as human-computer interaction, intelligent healthcare, safe driving, and social media analysis. In recent years, with breakthroughs in deep learning, CNNs have become the dominant approach in this field[1]. Compared to traditional methods that rely on handcrafted features,e.g.Local Binary Patterns (LBP), Histogram of Oriented Gradients(HOG) and machine learning classifiers,e.g. Support Vector Machine (SVM),

*Corresponding author's email: chenyitong916@gmail.com

CNNs leverage powerful end-to-end learning capabilities[2] and hierarchical feature abstraction mechanisms[3]. These allow the model to extract features ranging from basic edges and textures to complex facial organ combinations and expression patterns, providing strong support for accurate expression recognition. This has greatly advanced the field, significantly improving both the accuracy and efficiency of facial expression recognition.

While CNNs have demonstrated significant potential in FER, their deployment in real-world applications presents several persistent challenges, which have consequently shaped the primary focus of recent research. A central line of inquiry involves enhancing model robustness in unconstrained environments. Researchers have actively developed algorithms to effectively address significant variations in illumination, head pose, and partial occlusions (e.g., masks, glasses), aiming to bridge the performance gap between controlled lab datasets and complex real-world scenarios. Furthermore, the recognition of subtle and transient expressions, such as micro-expressions, has driven the design of network architectures capable of capturing fine-grained spatial-temporal features with greater sensitivity. Another critical research thrust addresses mitigating data bias and improving model generalization across diverse populations, necessitating the learning of more discriminative and invariant features that account for cultural, racial, and individual disparities. Finally, driven by the needs of mobile and embedded applications, there is a growing effort to strike a balance between model accuracy and computational efficiency, leading to the exploration of various lightweight network solutions. These collective challenges have steered the CNN-based FER field toward developing more robust, precise, fair, and practical systems.

To address the above challenges, this review comprehensively summarizes recent research advances in CNN-based facial expression recognition. It provides an in-depth analysis of the strengths and limitations of various network architectures in recognizing expressions, and summarizes the performance of different networks across various expression types and imaging conditions. This paper focuses specifically on specialized architectures and strategies designed to enhance model robustness, improve micro-expression recognition, increase interpretability, and facilitate better generalization. These include, for example, the incorporation of attention mechanisms to focus on key facial regions, the use of generative adversarial networks for data augmentation or domain adaptation, and the exploration of dual-path architectures that integrate both global and local features. By synthesizing, comparing, and analyzing these studies, this review aims to provide a clear picture of the current technological landscape, clarify the advantages and suitable scenarios of different technical routes, and point out promising future research directions. It is intended to serve as a comprehensive and valuable reference for subsequent research on CNN-based facial expression recognition.

2 Method

The evolution of FER has been significantly driven by innovations in CNN architectures. To address the inherent challenges such as occlusion, pose variation, and the subtlety of expressions, researchers have moved beyond standard CNN models towards more specialized designs. These methods can be broadly categorized into several key paradigms: attention-based mechanisms, multi-modal fusion, and the integration of local-global features.

2.1 Attention-based CNN Architectures

Attention mechanisms, inspired by human visual perception, allow networks to dynamically focus on the most discriminative regions of a face (e.g., eyes, mouth), while suppressing irrelevant or noisy information. This has proven highly effective in improving robustness.

2.1.1 Channel Attention

Channel attention modules aim to recalibrate the importance of different feature maps (channels) generated by a CNN. Each feature map can be considered a specific "detector" for a particular pattern. Channel attention learns to emphasize feature maps that are crucial for expression recognition and suppress less useful ones. The Squeeze-and-Excitation (SE) Network [4] is a pioneering example. It incorporates a lightweight module that first squeezes global spatial information from each feature map into a channel descriptor, then excites (reweights) the channels based on their learned importance. In FER, SE blocks can be integrated into backbone networks like ResNet, enabling the model to prioritize feature channels corresponding to expressive facial components, thereby enhancing performance, especially under challenging conditions [5].

2.1.2 Spatial Attention

While channel attention focuses on "what" to look for, spatial attention determines "where" to look. It generates an attention map that highlights salient spatial regions within the feature maps. The Convolutional Block Attention Module (CBAM) [6] sequentially applies both channel and spatial attention. The spatial attention component typically uses max-pooling and average-pooling operations across channels to create a spatial attention map, which is then multiplied with the input features. This forces the network to concentrate on semantically rich areas like furrowed brows or smiles, making it more resilient to occlusions and complex backgrounds. For instance, Wang, K., et al.[7] employed a spatial attention mechanism guided by facial landmarks to ensure the network focuses on predefined expression-relevant regions.

2.2 Multi-modal Fusion Methods

To overcome the limitations of relying solely on RGB intensity information, multi-modal CNN approaches fuse complementary data from different sources. This provides a more comprehensive representation of facial expressions.

2.2.1 Geometric Feature Fusion

Facial expressions are characterized by the deformation of facial geometry (e.g., the movement of key points). Multi-modal networks often fuse appearance features from the RGB image with geometric features extracted from facial landmarks. A common strategy is a two-stream network. One CNN stream processes the raw face image, while another stream, often a simpler network or even a fully connected network, processes the coordinates of facial landmark points. The features from both streams are then fused at a middle or late stage of the network. Liu, Y., et al. [8] demonstrated that such fusion significantly improves recognition accuracy by explicitly encoding the structural changes of the face that are less affected by illumination changes compared to pure appearance.

2.2.2 Multi-modal Fusion for Occlusion Handling

Occlusions severely degrade FER performance by removing critical facial features from RGB images. Multi-modal fusion enhances robustness by integrating complementary data. Thermal imaging captures heat patterns; since materials like fabric masks may be thermally transparent, underlying expression-related physiological signals remain detectable. Depth data provides 3D facial geometry, which is invariant to appearance changes and offers

discriminative cues from unobstructed areas. Advanced strategies employ attention mechanisms to dynamically weight modalities, suppressing occluded RGB regions while emphasizing reliable thermal/depth features. Generative models can also reconstruct occluded appearances using information from intact modalities. Despite their effectiveness, challenges include the cost of multi-modal data acquisition and increased model complexity.

2.3 Integration of Local and Global Features

Facial expressions involve both local muscle movements and their coordination across the entire face. Architectures that explicitly model this hierarchy have shown superior performance.

2.3.1 Multi-pathway and Region-based Networks

These networks explicitly design separate pathways for processing local facial regions and the global face.A typical architecture might consist of a global pathway that takes the entire face as input, and several local pathways that take cropped regions-of-interest (ROIs) around crucial parts like the eyes, nose, and mouth. The features from all pathways are then combined for the final classification. Zhi, R., et al.[9] proposed a Deep Region and Multi-label Learning (DRML) network, which uses a single CNN but employs a local convolutional layer to automatically learn region-specific features. This design ensures that local subtle cues are preserved and effectively integrated with global contextual information, which is critical for distinguishing between similar expressions (e.g., surprise and fear).

2.3.2 Dynamic Feature Extraction with 3D CNNs

While most methods focus on static images, expressions are inherently dynamic processes. 3D CNNs extend the concept of convolution into the temporal dimension, allowing for the direct extraction of spatio-temporal features from image sequences or videos. Models like 3D ResNet [10] and I3D [11] have been adapted for FER. By processing a short sequence of frames, these networks can capture the temporal evolution of an expression, such as the gradual onset of a smile or the rapid twitch of a micro-expression. This provides a richer feature set that is often more discriminative than static features alone.

3 Discussion

Despite the significant advances in CNN-based FER, the journey towards creating systems that operate as reliably and nuancedly as humans in real-world settings is far from over. This section synthesizes the core challenges that persist and outlines promising avenues for future research.

3.1 Persistent Challenges

3.1.1 Data Bias and Lack of Generalization

The performance of deep learning models is intrinsically tied to the distribution of their training data. A major hurdle for FER is the pervasive bias in existing large-scale datasets, which often overrepresent certain demographics (e.g., age, ethnicity) and expression intensities (e.g., exaggerated prototypical expressions) [12]. Consequently, models trained

on these datasets suffer from poor generalization when deployed in diverse cultural contexts or on populations not well-represented in the training data. For instance, expressions and their interpretations can vary significantly across cultures, leading to misclassification [13]. Achieving true algorithmic fairness and robustness requires a fundamental shift towards more inclusive and representative data collection.

3.1.2 Robustness in Unconstrained Environments

While methods like attention mechanisms and multi-modal fusion have improved robustness, FER systems still struggle with the immense variability of the real world. Challenges such as extreme and non-uniform illumination, significant head pose variations, and partial occlusions (e.g., masks, hands, accessories) remain difficult to fully mitigate. Most models are trained and evaluated on lab-controlled, front-facing datasets, creating a significant simulation-to-reality gap. Developing models that are invariant to these “non-expression” factors is crucial for practical applications like security monitoring or in-car driver monitoring systems.

3.1.3 The Complexity of Micro-expression Recognition

Micro-expressions are brief, involuntary facial movements that reveal concealed emotions. Their short duration (typically 1/25 to 1/5 of a second) and low intensity make them exceptionally challenging to capture and classify [14]. Standard CNNs, designed for spatial feature extraction, are often inadequate for analyzing these subtle spatiotemporal signals. Although 3D CNNs and video-based approaches offer a path forward, they require high-frame-rate data and are computationally intensive. The limited size and diversity of available micro-expression datasets further exacerbate this challenge.

3.1.4 Computational Efficiency and Real-time Deployment

Many state-of-the-art FER models, incorporating complex attention modules or multi-pathway architectures, have high computational demands. This poses a significant barrier for deployment on resource-constrained devices, such as mobile phones, embedded systems in vehicles, or edge computing nodes, where low latency and power efficiency are paramount. There is a pressing need to balance model accuracy with computational cost, often requiring specialized model compression, quantization, or the design of inherently efficient architectures tailored for FER.

3.2 Future Prospects

3.2.1 Neuro-Symbolic Integration and Explainable Model Building

To overcome data bias and enhance model trustworthiness, future research must steer FER from "black-box" correlation learning towards "white-box" causal reasoning. A core pathway is neuro-symbolic integration, which involves embedding prior knowledge, such as the Facial Action Coding System (FACS), into deep learning frameworks. By constructing differentiable symbolic reasoning layers, models can explicitly detect action units and learn the causal relationships between them and expressions [15]. This not only makes the model's decision-making process transparent, ensuring it relies on genuine muscle movements rather than spurious background correlations, but also lays the foundation for trustworthy applications in high-stakes domains like healthcare and security.

3.2.2 Practical and Efficient Dynamic Inference Frameworks

To resolve the conflict between model complexity and the resource constraints of embedded devices, future research on lightweight models must move beyond static model compression towards dynamic adaptive inference. Such frameworks allow the model to select different computational paths in real-time based on the complexity of the input sample (e.g., level of occlusion, image quality). For instance, a lightweight sub-network could be activated for clear, frontal faces, while more complex attention or temporal analysis modules are engaged for challenging samples with occlusions or suspected micro-expressions. This "on-demand" computational paradigm can significantly improve the energy efficiency and response speed of end devices while maintaining high average accuracy, which is key to the widespread deployment of FER.

3.2.3 Privacy-Preserving Data Collaboration via Federated Learning

The root causes of data bias and scarcity are data silos and privacy constraints. Federated Learning (FL) offers a highly promising solution, enabling multiple participants to collaboratively train a model on their local data without sharing the raw data. By establishing a federated data ecosystem spanning diverse cultures and scenarios, it is possible to train global models with stronger generalization capabilities and less bias, while strictly protecting the privacy of users' facial biometric information. This represents a systematic approach to addressing generalization and fairness challenges at the data source.

3.2.4 Continual Learning for Adaptation in Open Environments

The real physical world and application scenarios are constantly evolving. A practical FER system must be capable of continual learning, acquiring new knowledge (e.g., new types of occlusions, new cultural expression patterns) while avoiding catastrophic forgetting of previously learned knowledge. The research focus will be on developing continual learning algorithms suitable for FER, enabling models to perform efficient incremental updates with small amounts of new data and to recognize unknown expressions or out-of-distribution samples in an open world. This is central to ensuring the long-term stability and practicality of FER systems.

4 Conclusion

This review has comprehensively examined the advancements in CNNs applied to FER. It initially establishes the marked superiority of CNNs over traditional methods in feature learning and recognition performance, while also delineating the core challenges encountered when dealing with real-world complexities. In response, the research community has developed a range of effective solutions. This article has systematically organized and critically analyzed three predominant technical pathways: the introduction of attention mechanisms to guide models towards discriminative regions; the fusion of multi-modal data to leverage complementary information, thereby enhancing robustness against occlusions and illumination variations; and the design of architectures that integrate local and global features, including the use of 3D CNNs to capture spatio-temporal dynamics for more refined expression pattern analysis. Nevertheless, current research exhibits notable limitations. Inherent biases in datasets constrain model generalizability and fairness; model robustness in unconstrained environments requires further strengthening; for micro-expression recognition, existing methods struggle to balance sensitivity with computational efficiency; and the real-time deployment of complex models remains challenging. To address these issues, future

research should focus on the following directions: promoting neuro-symbolic integration to construct interpretable, causality-driven models; developing dynamic inference frameworks to balance accuracy and efficiency; adopting the federated learning paradigm to aggregate diverse data while preserving privacy, thereby enhancing generalizability at its source; and exploring continual learning mechanisms to enable model adaptation within evolving, open-ended environments. In summary, sustained exploration along these avenues promises to steer CNN-based FER technology toward more robust, equitable, efficient, and practical development, ultimately enabling it to deliver its full potential across a spectrum of real-world applications.

References

1. A.Mollahosseini, D.Chan, M.H.Mahoor, Going deeper in facial expression recognition using deep neural networks, in Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), 2019.
2. A.Krizhevsky, I.Sutskever, G.E.Hinton, ImageNet classification with deep convolutional neural networks, in Advances in Neural Information Processing Systems, vol. 25, 2012.
3. Y.LeCun, L.Bottou, Y.Bengio, P.Haffner, Gradient-based learning applied to document recognition, in Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, 1998.
4. J.Hu, L.Shen, G.Sun, Squeeze-and-excitation networks, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.
5. Y.Li, J.Zeng, S.Shan, X.Geng, A Squeeze-and-Excitation based Deep Network for Facial Expression Recognition, Pattern Recognition Letters, 2020.
6. S.Woo, J.Park, J.-Y.Lee, I.S.Kweon, Cbam: Convolutional block attention module, in Proceedings of the European Conference on Computer Vision (ECCV), 2018.
7. K.Wang, X.Peng, J.Yang, D.Meng, Facial Expression Recognition with Spatial Attention and Label Self-Adaption, IEEE Transactions on Affective Computing, 2020.
8. Y.Liu, H.Yuan, X.Liu, Geometry-Guided Multi-modal Fusion for Facial Expression Recognition, in Proceedings of the ACM International Conference on Multimedia, 2021.
9. R.Zhi, M.Liu, D.Zhang, DRML: Deep Region and Multi-label Learning for Facial Expression Recognition, IEEE Transactions on Image Processing, 2018.
10. K.Hara, H.Kataoka, Y.Satoh, Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.
11. J.Carreira, A.Zisserman, Quo vadis, action recognition? A new model and the kinetics dataset, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.
12. L.Rhue, Racial Influence on Automated Perceptions of Emotions, SSRN Scholarly Paper, 2018.
13. R.E.Jack, C.Blagrove, P.G.Schyns, Facial expressions of emotion are not culturally universal, Proceedings of the National Academy of Sciences, 2012.
14. S.T.Liong, Y.S.Gan, J.See, H.-K.Phan, Shallow CNN with Depthwise Separable Convolutions for Micro-Expression Recognition, IEEE Transactions on Affective Computing, 2019.
15. W.Li, F.Liu, J.Zhu, Knowledge-Guided Deep Facial Expression Recognition, IEEE Transactions on Affective Computing, 2021.