Vol. 20 (2026): Proceedings of 2026 International Conference on Materials Science, Energy Engineering and Automation Control (ICMSEAC 2026)
E-ISSN: 3105-0522 P-ISSN: 3105-0514
DOI: https://doi.org/10.70267/mseac.20262025

Technology for Facial Expression Recognition Across Different Scenes

Hongrui Zhang*

School of Machinery and Automation, Wuhan University of Science and Technology, Wuhan, China

**Corresponding author: Hongrui Zhang.*

Abstract

This paper conducts a systematic review and analysis of the key technologies for facial expression recognition across different scenarios. As the facial expression recognition technology moves from the laboratory to diverse real-world scenarios such as remote healthcare, mental health, and intelligent cabins, ensuring the model maintains high performance in complex environments has become a core issue that needs to be urgently addressed. Firstly, this paper reviews three key technology systems focusing on environmental robustness, efficient deployment, and data scarcity and domain adaptation. It elaborates on representative methods such as attention mechanisms, lightweight networks, and domain adaptation, along with their progress. Based on this, it further analyzes the common challenges faced in cross-scenario applications, including environmental interference, data distribution differences, and deployment constraints. The review results indicate that there is no universal optimal model. The technical solutions need to be deeply adapted to specific scene requirements. The integration of robustness, lightweighting, and domain adaptation technologies is the future development trend. This paper provides theoretical basis and practical references for the selection of facial expression recognition technologies for specific scenarios.

Keywords

facial expression recognition, environmental robustness, lightweight, domain adaptation, cross-scene

1. Introduction

Facial expression recognition is one of the core technologies in affective computing and human-computer interaction. In recent years, it has made significant progress in controlled environments. However, when the technology is applied to real scenarios such as remote surgery, mental health assessment, and intelligent cockpits, factors such as changes in lighting, differences in head postures, and facial occlusion lead to a sharp decline in model performance. Different scenarios have fundamentally different requirements for the technology: the surgical scenario demands high reliability and real-time performance, clinical diagnosis emphasizes sensitivity and explainability to subtle expressions, and human-computer interaction pursues lightweight deployment and cross-individual generalization capabilities. Currently, although the academic community has developed various advanced algorithms, there is a lack of systematic review of the technology system from a cross-scenario perspective. This paper aims to fill this gap, summarize and review

the key technologies for cross-scenario requirements, analyze their characteristics and applicable boundaries, and provide references for technology selection and application.

2. The Key Technical Framework for Addressing Cross-scenario Challenges

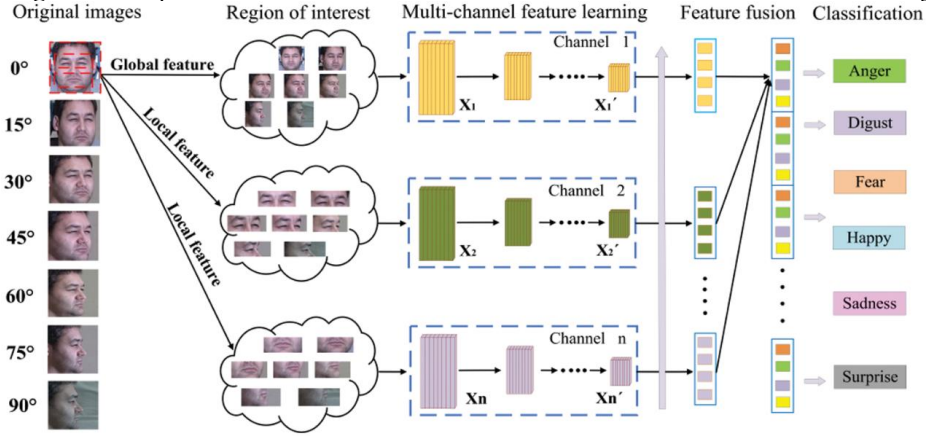
2.1 Technologies for Environmental Robustness

The variations in lighting conditions, differences in head postures, and facial occlusions in real scenes generally pose an interference to expression recognition. To address these environmental robustness challenges, researchers have proposed various solutions from different perspectives.

In the aspect of attention mechanism, Liu et al. proposed an adaptive multi-layer perceptual attention network, which designs a cascaded structure of channel attention and spatial attention to guide the model to adaptively focus on the information-rich facial regions, effectively suppressing background noise interference [1]. This network has been verified on multiple benchmark datasets for its robustness against complex backgrounds.

In terms of pose invariance, Liu et al. proposed a deep global multi-scale and local block attention dual-branch network [2], whose overall architecture is shown in Figure 1.

Figure 1: Deep Global Multi-scale and Local Block Attention Dual-Branch Network [2]



The core innovation of this network lies in the dual-branch collaborative design: the upper branch utilizes the self-attention mechanism to extract the global multi-scale features of the entire face. The self-attention mechanism calculates the correlation between any two positions in the feature map, enabling the capture of long-distance context dependencies, thereby automatically locating key areas such as eyes and mouths in the global view; the lower branch divides the face into multiple local blocks (such as the left eye, right eye, nose, mouth, etc.), extracts local features through the relationship attention module, and models the spatial correlations between these local blocks - for example, both the left and right eyes squinting together constitute a genuine smile, while a single eye change may be an accidental action. Finally, the feature fusion layer adaptively weights and fuses the global context and local detail information, and inputs them into the classifier for expression recognition. This design enables the model to provide reliable cues even when the entire face information is incomplete due to large head rotations, significantly improving the recognition performance under posture change conditions.

Regarding the robustness against occlusion, Huang et al. proposed FERMixNet, which combines the facial mixture enhancement strategy with mid-level representation learning [3]. This model randomly selects two face images during the training stage, generates new training samples through region mixing to simulate local occlusion situations in real scenarios; at the same time, it introduces a mid-level representation learning module to force the network's intermediate layers to learn invariant feature representations of occlusion, enabling the model to maintain stable recognition even when encountering unseen occlusion types during testing. Ma et al. designed the FER-VMamba framework, introducing a state space model [4]. Through global compact attention and hierarchical feature interaction mechanisms, it achieves long-distance feature

dependency modeling while maintaining linear computational complexity, demonstrating excellent robustness in complex lighting and occlusion environments.

These studies show that the combination of the attention mechanism, multi-scale feature fusion, and occlusion modeling techniques is an effective technical path for improving environmental robustness.

2.2 Lightweight Technologies for Efficient Deployment

In practical applications, the facial expression recognition models need to run on resource-constrained embedded devices, meeting engineering constraints such as real-time performance and low power consumption. To address this requirement, lightweight technologies have become a research hotspot, mainly focusing on three aspects: network structure design, model compression, and edge computing architecture.

In the aspect of lightweight network design, Zhao et al. proposed a lightweight facial expression recognition model for automatic engagement detection, replacing standard convolution with depthwise separable convolution, which reduces the parameter quantity to less than 0.5M while maintaining high recognition accuracy, making it suitable for deployment on mobile devices [5]. Mou et al. proposed a lightweight method based on class-balanced fusion cumulative learning, addressing the common class imbalance problem in facial expression recognition, and designed an adaptive weighted loss function to achieve balanced learning of various facial expressions in the lightweight network [6].

In the aspect of edge deployment, Yang et al. proposed a real-time facial expression recognition scheme based on edge computing [7]. This scheme adopts a three-layer edge computing architecture: the device layer is responsible for data collection by terminal devices such as vehicle-mounted cameras; the edge layer deploys a lightweight model to process data in real time on the vehicle chip, achieving a latency of less than 50ms for fatigue state monitoring; the cloud layer is responsible for data aggregation and model iteration updates. The advantage of this architecture lies in real-time response, privacy protection, and the elimination of network dependence. Riaz et al. designed the TriViT-Lite model, which combines the advantages of Vision Transformer and MobileNet, achieving a balance between the global modeling ability of Transformer and the local inductive bias of CNN through a texture-aware attention mechanism, while compressing the model parameters to 3.2M while maintaining a recognition accuracy of 91.2% [8].

2.3 Technologies for Dealing with Data Scarcity and Domain Adaptation

The emotion recognition model highly relies on large-scale labeled data, but the significant differences in data distribution among different scenarios severely restrict the generalization ability of the model. Li and Deng conducted a systematic analysis of the bias in facial expression datasets and revealed the distribution differences among different datasets through large-scale cross-database experiments, pointing out that the laboratory performance-based expressions and the real spontaneous expressions have essential differences in their presentation methods [9].

To address this challenge, domain adaptation technology has become the focus of research. Yan et al. proposed a multi-feature representation multi-layer domain adaptation fusion network that adopts a multi-level domain alignment strategy [10]. Through multi-layer domain discriminators conducting adversarial training at different feature levels, it narrows the distribution gap between the source domain and the target domain at multiple abstract levels. Kim et al. proposed the DIRE framework, which explicitly decouples learning and decomposes features into domain-related components and domain-independent components [11]. By imposing orthogonal constraints to force the two to be independent, it trains the model to perform expression classification only using the domain-independent components, thereby ensuring the model's generalization ability on unseen target domains. Chen et al. constructed a unified evaluation benchmark for cross-domain emotion recognition and proposed a cross-domain recognition method based on adversarial graph learning [12]. Through graph convolution networks transmitting domain adaptation information in adversarial training, it provides important references for fair comparisons and algorithm evaluation within the domain.

3. Core Challenges in Cross-Scene Applications

Although the techniques discussed in Chapter 2 laid a solid foundation for cross-scenario facial expression recognition, in practical deployment, this technology still faces a series of common and critical

problems. These problems can be summarized into three main aspects: environmental interference, data distribution differences, and deployment limitations. Below, each type of challenge will be analyzed, and at the end of each section, the "Corresponding Technical Solution Analysis" will be provided, clearly indicating the type of technology that is most suitable for solving the challenge and its connection to the literature.

The most direct interference comes from the uncontrolled environment. In laboratory conditions, the lighting is uniform, the head posture is fixed (usually in a frontal position), and there are no obstructions. However, in real scenarios such as intelligent cabins and operating rooms, the lighting conditions vary greatly - from strong backlighting, weak light to dynamic changes when entering or exiting tunnels; the head posture freely rotates with the user's turning, lowering the head, etc.; masks, glasses, hair, etc. are frequently obstructive. These factors together cause the facial features extracted by the model to be severely distorted, resulting in a sharp decline in recognition performance. Studies have shown that models with an accuracy rate of over 95% on controlled datasets may drop to below 60% when encountering large-angle head deviations or partial obstructions. To address the challenge of environmental robustness, the most effective technical approach is the attention mechanism combined with multi-scale feature fusion. The self-attention module in the dual-branch network proposed by Liu et al. can automatically focus on key facial areas such as eyes and mouth even under pose variations, while the relation attention module explicitly divides the face into multiple local blocks and uses local details for compensation when global information is lacking [2]. Similarly, FERMixNet adopts a facial mixture enhancement strategy, simulating occlusion situations during the training phase and forcing the model to learn invariant mid-level representations under occlusion [3]. The FER-VMamba framework introduces global compact attention and hierarchical feature interaction mechanisms, maintaining robustness in complex lighting and partial occlusion environments [4]. The common principle of these methods is that when global features are disrupted, the model can rely on reliable local cues. Therefore, for unpredictable scenarios such as driver monitoring or remote surgery (e.g., where lighting, pose, or occlusion is unknown), deploying a robust model based on the attention mechanism and enhanced training is the most appropriate strategy.

The second major challenge lies in the issue of data acquisition and distribution disparity. Obtaining high-quality labeled facial expression data is extremely costly, especially for micro-expressions which require expert frame-by-frame annotation. Moreover, there is a significant domain shift between different databases. For instance, the performance-based expression datasets collected in laboratories (such as CK+) differ greatly from those collected online (such as AffectNet) in terms of presentation. When models trained on CK+ are transferred to AffectNet, their accuracy often drops significantly. Additionally, factors such as cultural background, age, and gender can also cause individual differences in facial expressions, further increasing the difficulty for the models to generalize across different user groups. The most suitable techniques for alleviating data and generalization challenges are domain adaptation and domain-invariant representation learning. Yan et al. proposed a multi-feature representation multi-layer domain adaptation fusion network, which aligns the feature distributions of the source domain and the target domain at multiple levels through adversarial training [10]. Kim et al. proposed the DIRE framework, which explicitly decouples features into domain-dependent and domain-independent parts, and only uses domain-independent features for classification, enabling the model to generalize to unseen target domains [11]. Chen et al. constructed a unified evaluation benchmark for cross-domain expression recognition and proposed a domain adaptation method based on adversarial graph learning, which transmits domain alignment information along the similarity graph using graph convolution networks [12]. These methods all aim to solve the core problem of "domain shift", forcing the model to learn common representations between different data sources. Therefore, for scenarios with scarce target domain labeled data such as clinical diagnosis or cross-cultural applications, domain adaptation and domain-invariant learning are the preferred technical paths.

Finally, deployment limitations represent the "last mile" for technology implementation. The computing power, memory, and power consumption of edge devices (such as vehicle chips, mobile phones, and embedded sensors) are limited, which creates a sharp contradiction with the high computational requirements of deep learning models. At the same time, interaction scenarios have strict requirements for real-time performance (for example, driver fatigue monitoring needs to provide results within 50 milliseconds) and privacy security (face images should not leave the local device). These constraints often force designers to make trade-offs between accuracy and efficiency. The most effective solution to address deployment constraints is lightweight network

design and edge computing architecture. Zhao et al. proposed a lightweight model for automatic participation degree detection, which reduced the parameter quantity to below 0.5M while maintaining a high accuracy rate [5]. Mou et al. proposed a method based on class-balanced fusion cumulative learning, which solved the problem of class imbalance in lightweight networks [6]. Yang et al. proposed a real-time facial expression recognition scheme based on edge computing, deploying the compressed model on the vehicle chip to achieve a delay of less than 50ms, without relying on the cloud [7]. Riaz et al. designed the TriViT-Lite model, which integrates Vision Transformer and MobileNet, achieving an accuracy rate of 91.2% with 3.2M parameters, and is suitable for resource-constrained medical devices [8]. These technologies directly address hardware limitations by reducing model size and enabling local processing. Therefore, for edge interaction applications such as intelligent cabins and online education, the combination of lightweight models and edge computing is the most suitable technical choice.

The above three types of challenges are not independent of each other in practical applications; instead, they often interweave and restrict each other. For instance, the attention mechanism or multi-scale fusion models designed to enhance environmental robustness often come with a significant increase in parameters, which directly contradicts the lightweight requirements in deployment constraints - the more robust the model is, the more difficult it is to run in real-time on edge devices. Another example is that although domain adaptation methods can effectively alleviate data distribution differences, their training process usually requires simultaneous access to source domain and target domain data, resulting in high computational costs and posing pressure on edge deployment. Conversely, excessive pursuit of lightweighting may lead to a decline in feature extraction capabilities, thereby weakening the model's resistance to environmental disturbances and domain shifts.

This inherent interdependence indicates that there is no single universal model that can simultaneously perfectly meet all requirements. The design of any practical system must make trade-offs and selections based on the core requirements of the target application scenario.

4. Conclusions

This article systematically reviews the key technical systems for cross-scenario facial expression recognition, covering three core directions: environmental robustness, lightweight design, and domain adaptation. It also deeply analyzes the common challenges faced by cross-scenario applications. The research shows that there is no universal model that applies to all situations; the technical solutions must be deeply adapted to specific scene requirements: high-reliability scenarios should prioritize the combination of robustness and real-time performance, high-sensitivity scenarios should focus on domain adaptation and fine-grained feature extraction, and scenarios with high generalization requirements should combine lightweight design with domain adaptation technologies. Current technologies are showing a trend of integration, such as the parallel design of robustness and lightweight, the balance between interpretability and efficient deployment, and the combination of domain adaptation and online learning.

Looking to the future, dynamic continuous adaptation, privacy protection learning, the cognitive leap from emotion recognition to emotion understanding, and the establishment of a standardized evaluation system will be important directions worthy of continuous exploration. With the development of new human-machine interaction carriers such as soft robots, combining facial expression perception with compliant execution is expected to achieve truly “warm” intelligent companions.

References

[1] Liu, H., Cai, H., Li, Q., et al. (2022). Adaptive Multilayer Perceptual Attention Network for Facial Expression Recognition. *IEEE Transactions on Circuits and Systems for Video Technology, 32*(9), 6253-6266.

[2] Liu, C., Liu, X., Chen, C., et al. (2024). Deep Global Multiple-Scale and Local Patches Attention Dual-Branch Network for Pose-Invariant Facial Expression Recognition. *Computer Modeling in Engineering & Sciences, 139*(1), 405-440.

[3] Huang, Y., Peng, J., Zhang, W., et al. (2025). FERMixNet: An Occlusion Robust Facial Expression Recognition Model With Facial Mixing Augmentation and Mid-Level Representation Learning. *IEEE Transactions on Affective Computing, 16*(2), 639-654.

[4] Ma, H., Lei, S., Li, H. C., et al. (2025). FER-VMamba: A robust facial expression recognition framework with global compact attention and hierarchical feature interaction. *Information Fusion, 124*, 103371.

[5] Zhao, Z., Li, Y., Yang, J., et al. (2024). A lightweight facial expression recognition model for automated engagement detection. *Signal, Image and Video Processing, 18*(4), 3553-3563.

[6] Mou, X., Song, Y., Wang, R., et al. (2023). Lightweight Facial Expression Recognition Based on Class-Rebalancing Fusion Cumulative Learning. *Applied Sciences, 13*(15), 9029.

[7] Yang, J., Qian, T., Zhang, F., et al. (2021). Real-Time Facial Expression Recognition Based on Edge Computing. *IEEE Access, 9*, 76178-76190.

[8] Riaz, W., Ji, J., & Ullah, A. (2025). TriViT-Lite: A Compact Vision Transformer–MobileNet Model with Texture-Aware Attention for Real-Time Facial Emotion Recognition in Healthcare. *Electronics, 14*(16), 3256.

[9] Li, S., & Deng, W. (2022). A Deeper Look at Facial Expression Dataset Bias. *IEEE Transactions on Affective Computing, 13*(2), 881-893.

[10] Yan, J., Du, C., Li, B., et al. (2025). Cross-database facial expression recognition based on Multi-feature Representation Multi-layer Domain Adaptive Fusion Network. *Engineering Applications of Artificial Intelligence, 155*, 110995.

[11] Kim, H., Jung, Y., & Song, B. C. (2025). DIRE: Enhancing Facial Expression Recognition Through Domain-Invariant Representation Learning for Robust Generalization. *IEEE Transactions on Multimedia, 27*, 9542-9554.

[12] Chen, T., Pu, T., Wu, H., et al. (2022). Cross-Domain Facial Expression Recognition: A Unified Evaluation Benchmark and Adversarial Graph Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(12), 9887-9903.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

Open Access