Dean&Francis
ISSN 2959-6157

Research and Application Analysis of Image-Based Sentiment Analysis Methods

Yuxuan Shi[1]

and Beichen Zhang[2,*]

[1]Department of Computer Convergence Software, Korea University, Sejong, Korea
[2]Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, China
*Corresponding author: 1155211153@link.cuhk.edu.hk

Abstract:

Emotion, as the core of human subjective experience, exerts a profound impact on mental health and social interactions. With the development of computer vision technology, image-based emotion analysis has become a critical research direction, finding wide application in driving safety, advertising marketing, healthcare, and other fields. However, the diversity and complexity of emotional expressions, along with environmental interference and challenges in multimodal fusion, still hinder the improvement of its accuracy and robustness. This paper provides a comprehensive review and analysis of image-based emotion analysis methods and their practical applications. First, it briefly introduces the research background and application value of this field, highlighting its key roles in scenarios such as driver/passenger emotion monitoring in autonomous driving and consumer emotion recognition in advertising. Subsequently, four categories of core technical methods are elaborated in detail. This paper also conducts a systematic comparison of the four method categories in terms of their core technical frameworks, key innovations, performance (efficiency and robustness), applicable scenarios, and limitations. Finally, it summarizes three major current challenges: the diversity of emotional expressions, data annotation issues, and multimodal fusion difficulties. Corresponding future directions are proposed, including enhancing model robustness and generalization, optimizing data and efficient models, and advancing adaptive multimodal fusion technologies, aiming to provide comprehensive technical references for the development and application of this field.

Keywords: Image-Based Emotion Analysis; Multimodal Fusion; Vision Transformer; Lightweight Models

ISSN 2959-6157

1. Introduction

Emotion, as the core of human feelings, is expressed in diverse and complex ways, influencing mental health and social interactions significantly. With the development of computer vision, image-based emotion analysis has become a critical research direction with wide application value.

In driving safety and autonomous vehicles, image-based emotion analysis plays an important role: it can not only assess drivers' emotional states to prevent hazards but also shift focus to passengers with the popularization of autonomous vehicles—for example, adjusting driving styles based on passengers' well-being, preventing safety risks caused by passengers' emotional abnormalities, and optimizing route planning using passengers' emotional data. However, this application also faces challenges such as clarifying emotion triggers, achieving passenger diarization in multi-person scenarios, and solving audio-visual occlusion issues [1].

In advertising marketing, images are key to ads evoking consumer emotions, and a single ad image often conveys multiple emotions. Accurately identifying these emotions helps improve ad efficiency (currently, only less than 50% of ads attract consumers' attention). Existing research uses graph convolutional neural networks combined with emotion label semantic relations to optimize emotion prediction for ad images, achieving 14% higher mean average precision and 4% higher overall F1 score than traditional models, supporting large-scale accurate ad delivery [2].

However, the diversity and complexity of emotional expressions (e.g., in-vehicle environmental interference for autonomous driving, skewed emotion label distribution in advertising) still pose major challenges to image-based emotion analysis, which needs further research to address.

Recent breakthroughs in deep learning have injected significant momentum into image-based emotion recognition, driving advancements across multiple technical fronts. In the integration of physiological signals and visual analysis, facial thermography approaches have achieved robust modeling of dynamic emotional arousal through Statistical Parametric Mapping (SPM) and Hemodynamic Response Function (HRF) applications, overcoming limitations of traditional local region analysis while ensuring non-contact privacy protection [3]. Sato et al. demonstrated that skin temperature variations in specific facial regions correlate strongly with emotional states, providing a physiological foundation for thermal imaging-based emotion detection [4].

In multimodal fusion research, Cheng et al. introduced Emotion-LLaMA, which integrates audio (HuBERT), visual (MAE), and textual features through emotion-specific encoders. By constructing the fine-grained annotated MERR dataset (28,618 coarse-grained and 4,487 fine-grained samples) and employing instruction tuning, their model achieved superior zero-shot generalization with a 0.9036 F1 score on the MER 2023-Semi challenge [5].

Lightweight model development has seen notable progress through architecture optimization, with models like EmotionNet-X compressing parameters to 19.9M while maintaining 99.86% accuracy on the CK+ dataset, enabling efficient edge deployment with 18ms inference time [6]. Multitask learning frameworks, such as UFEN+MTFN, leverage Bi-GRU and multi-scale CNNs with cross-attention mechanisms to facilitate cross-modal interactions, enhancing robustness in complex environments with potential modality missing [7]. These advancements collectively expand the precision, efficiency, and applicability of image-based emotion recognition systems.

The aim of this paper is to provide a comprehensive review of image-based emotion analysis methods and their applications. The structure of this paper is as follows: The second section will provide a detailed introduction to the research background and development trajectory of image-based emotion analysis methods; The third section will thoroughly analyze the strengths and weaknesses of existing approaches and discuss the challenges currently faced by research; The fourth part will explore the application prospects of image-based emotion analysis technology across various fields; Finally, the fifth section will summarize the entire paper and propose future research directions.

2. Methods

This chapter outlines core methods for image-based emotion analysis, focusing on single-modal precise modeling, Vision Transformer (ViT) feature enhancement, multi-modal deep fusion, and vision-language large model reasoning.

2.1 Single-Modal Image Analysis Methods

Single-modal methods lay the foundation for image-based emotion analysis, focusing on extracting emotional cues from a single data type while balancing accuracy, efficiency, and scenario adaptability.

2.1.1 Pixel-Level Facial Thermal Image Analysis

Facial thermal imaging technology stands out for its non-contact, privacy-preserving advantages—critical for sensitive scenarios like medical care or elderly monitoring. Budu et al. proposed a pixel-level analysis framework to address the limitation of traditional region-of-interest (ROI) methods (which only focus on the nose tip, ignor-

ing global thermal patterns) [3]. The workflow includes three core steps:

(1) Data acquisition and preprocessing: A 640×480 resolution thermal camera captures facial temperature distribution, and custom software tracks glabella landmarks to eliminate motion artifacts caused by head movements;

(2) Spatial alignment via UV mapping: 2D thermal images are projected onto a 3D facial template using Blender, correcting perspective distortion from different shooting angles;

(3) Statistical parametric analysis: Statistical Parametric Mapping (SPM) is used for pixel-wise correlation between thermal signals and dynamic arousal ratings, while the hemodynamic response function (HRF) models the 3–10 second delay of blood flow changes [8], ensuring temporal synchronization between emotional stimulation and thermal response.

Experiments involved 21 participants (11 females, mean age 22.3 years) watching five emotion-evoking films (anger, sadness, neutral, contentment, amusement). Results showed significant negative correlations between nose tip/forehead temperature and arousal (Z=3.0–3.7), with high accuracy for basic emotions in a 23.5–24.5°C controlled environment [3].

2.1.2 Lightweight CNN Optimization

Convolutional Neural Networks (CNNs) excel at RGB image-based emotion recognition, but traditional models (e.g., VGG19 with 138M parameters) are too bulky for edge devices. Abbas et al. developed EmotionNet-X, a lightweight CNN optimized for emotion analysis, with three key designs:

(1) Streamlined feature extraction: 4 convolutional layers (Conv2D, filters: 64→64→128→256) capture multi-scale features (e.g., facial edge textures, facial contours), with 7 Dropout layers and 1 batch normalization layer to suppress overfitting and improve training stability;

(2) Parameter compression: Through streamlined model structure design (optimizing convolutional and fully connected layers), parameters are reduced to 19.9M (approximately 14.4% of VGG19);

(3) Efficient training: Adam optimizer (learning rate=1e-3) and categorical cross-entropy loss are used, with batch size 64 and 100 epochs for training.

On the CK+ dataset, EmotionNet-X achieved 99.86% accuracy—surpassing lightweight models like MobileNetV2 and outperforming ResNet50V2 in parameter efficiency—with around 18ms inference time (about 8.3× faster than VGG19). It also showed certain generalization ability on the FER2013 dataset (specific accuracy not provided in the original paper), making it suitable for edge devices like smart watches (real-time emotion monitoring) and smart access control (emotion-based access authorization) [6].

2.2 Improved Vision Transformer (ViT) Models

ViT excels at global feature modeling but lacks inner-patch information interaction and has limited feature scale diversity, leading to suboptimal performance in dense prediction tasks (e.g., object detection, semantic segmentation). The following methods integrate CNN's inductive bias (local feature extraction, multi-scale representation) to address this limitation.

2.2.1 Convolutional Multi-Scale Feature Interaction

Xia et al. proposed ViT-CoMer, a hybrid architecture fusing CNN multi-scale spatial features and Transformer global attention, tailored for dense prediction tasks. Its core innovations include:

(1) Multi-Receptive Field Feature Pyramid (MRFP): Composed of linear projection layers and depthwise separable convolutions with multi-scale receptive fields (e.g., 3×3, 5×5), it captures spatial features across different scales—supplementing ViT's lack of local detail extraction by expanding the range of spatial information captured;

(2) CNN-Transformer Bidirectional Fusion Interaction (CTI) Module: A bidirectional feature fusion mechanism that unifies CNN and Transformer features via multi-scale deformable self-attention. This module aligns multi-resolution features (e.g., CNN's 1/8, 1/16, 1/32 resolution features with ViT's 1/16 resolution features) and enhances the synergy between local spatial cues and global context, addressing ViT's non-hierarchical feature limitation [9].

On benchmark datasets for dense prediction: On COCO val2017 (object detection/instance segmentation), ViT-CoMer-L (with DINOv2 pre-training) achieves 64.3% AP (bounding box) under the Co-DETR framework, outperforming state-of-the-art ViT-Adapter-L by 1.7% AP without extra training data; on ADE20K val (semantic segmentation), it reaches 62.1% mIoU (with Mask2Former), comparable to vision-specific backbones like Swin-L. These performances make it suitable for scenarios such as autonomous driving (real-time object detection) and remote sensing (large-scale scene segmentation) [9].

2.2.2 Inductive Bias Fusion for ViT via Hybrid Architectures

Haruna et al. (2025) reviewed hybrid CNN-ViT architectures that incorporate CNN's inductive bias into ViT, focusing on two representative design strategies for enhancing local feature modeling:

(1) Hierarchical Inductive Bias Injection: Architectures like ViTAE and CvT integrate convolutional blocks into

ISSN 2959-6157

Transformer layers. For example, ViTAE adds spatial pyramid reduction modules and convolution blocks to ViT, extracting local multi-scale features (e.g., edge, texture) while retaining global attention; CvT replaces ViT's positional encoding with convolutional implicit spatial information, improving parameter efficiency for local feature learning;

(2) Efficient Feature Fusion with Inductive Bias: Lightweight hybrid designs (e.g., LeViT, Efficient Former) use CNN's convolutional stems to reduce spatial dimensions early, lowering Transformer's computational cost (e.g., LeViT uses CNN stages to shrink feature maps before Transformer processing, cutting FLOPs by ~40% vs. vanilla ViT). These architectures often combine self-supervised pre-training (e.g., masked image modeling) to optimize generalization [10].

Practically, such hybrid architectures balance performance and efficiency: For instance, CvT-W24 achieves 87.7% top-1 accuracy on ImageNet-1k with fewer parameters than deep CNNs like ResNet-152; ViTAE-S outperforms T2T-ViT-14 by 1.5% top-1 accuracy on ImageNet while maintaining moderate computational cost. Their versatility makes them applicable to medical image analysis (e.g., organ segmentation) and industrial quality inspection (e.g., defect detection in manufacturing) [10].

2.3 Multi-Modal Fusion Models

Human emotions are inherently multi-modal (e.g., facial expressions + vocal tones + text). Multi-modal fusion methods integrate multi-source information to improve robustness.

2.3.1 Instruction-Tuned Cross-Modal Alignment

Cheng et al. developed Emotion-LLaMA, a multi-modal large language model optimized for emotion recognition, with three core designs:

(1) MERR dataset construction: 28,618 coarse-grained (basic emotions) and 4,487 fine-grained samples, with annotations covering audio (HuBERT as the audio encoder to extract vocal tone features), visual (MAE as the local encoder to extract static facial expression features), and text (LLaMA-3 refines semantic cues by aggregating unimodal descriptions);

(2) Cross-modal alignment: Cross-modal alignment: A trainable linear layer maps audio/visual features (4096D) to LLaMA's text embedding space (4096D), forming unified multi-modal tokens;

(3) Emotion-specific instruction tuning: Two core tasks (classification: "predict emotion labels from multi-modal inputs"; reasoning: "explain emotion-related cues across modalities") guide the model to handle cross-modal inconsistency (e.g., smiling face + angry tone) [5].

On the MER2023-Semi challenge, Emotion-LLaMA achieved an F1 score of 0.9036. In education, it can analyze students' classroom videos (visual: yawns; audio: low voice; text: "I get it") to identify hidden boredom; in customer service, it recognizes user frustration from "polite text + sharp tone" to adjust service strategies.

2.3.2 Hierarchical Fusion with Multi-Task Learning

Cai et al. proposed a framework combining Unimodal Feature Extraction Network (UFEN) and Multi-Task Fusion Network (MTFN):

(1) UFEN: For images, 2 CNN layers + Bi-GRU capture spatial-temporal features (e.g., smile duration); for audio, 1D CNN extracts 22050Hz acoustic features (MFCC, log F0); for text, Chinese-BERT extracts semantic features (dropout rate=0.1);

(2) MTFN: Low-level cross-modal attention (4 heads) integrates image-audio features (e.g., increasing audio weight in noisy environments); middle-level Transformer encoder-decoder reconstructs fused features; high-level concatenation for classification, with multi-task loss (single-modal emotion classification loss + multi-modal fusion loss + encoder-decoder feature reconstruction loss) optimizing parameters [7].

On the CMU-MOSI dataset, this framework achieved an MAE of 0.728 (lower than TFN); on the Chinese SIMS dataset, its 5-class accuracy reached 45.20% (higher than Graph-MFN). It shows strong robustness to modal imbalance—e.g., when partial visual features are missing, audio-text fusion still maintains stable performance.

2.4 Vision-Language Large Model Reasoning

Nadeem et al. proposed a comparative framework of vision-enabled large language models (LLMs) and fine-tuned deep learning models (DLMs) for image-based emotion recognition, with three key components:

(1) Visual feature extraction: DLMs (CNN_1, CNN_2, ResNet50, VGG16) use convolutional/residual structures (e.g., CNN_2 with 10 convolutional layers), while vision-LLMs (LLaVa, GPT-4) integrate visual-language capabilities (LLaVa combines CLIP's visual encoding and Vicuna's language processing);

(2) Cross-model evaluation alignment: A unified 560-image FER2013 test set (80 per emotion class, grayscale) ensures fair performance comparison;

(3) Interpretable reasoning: GPT-4 automatically explains emotion classification (e.g., „Happy"/"Sad" reasoning), while LLaV6a needs prompting for details, improving decision transparency.

On the FER2013 dataset, fine-tuned CNN_2 achieved the highest accuracy (62%), followed by GPT-4 (55.81%); LLaVa outperformed CNN_1 (48.21%) and VGG16

(31.01%), with consistent precision/recall/F1 trends. Limitations: Vision-LLMs lagged behind fine-tuned DLMs (e.g., GPT-4 < CNN_2); the small grayscale test set limits generalization; VGG16 (31.01% accuracy) is unstable for complex emotions. It suits small-dataset scenarios (e.g., scarce labeled emotion data) or no-additional-training tasks (e.g., rapid image-based emotion prototyping) [11].

3. Comparison and Discussion

3.1 Strenths and Limitations of Different Models

3.1.1 Pixel-Based Facial Thermal Imaging for Emotion Analysis

In terms of the core technical framework, this method follows a data flow of „thermal imaging data collection → facial landmark tracking via custom software → mapping of 2D thermal images to 3D facial models → pixel-level analysis using Statistical Parametric Mapping (SPM) → dynamic correction with Hemodynamic Response Function (HRF)". Its key technologies include SPM, Random Field Theory (RFT), and HRF, while inter-frame feature alignment is achieved through 3D facial models. The key innovations lie in replacing traditional Region of Interest (ROI) analysis with pixel-level full-face analysis to avoid the loss of facial edge information; using SPM + RFT to correct false positives caused by multiple comparisons, ensuring the statistical robustness of results; and integrating HRF to simulate the hemodynamic delay induced by emotional stimuli, thereby improving the accuracy of dynamic emotion capture.

In terms of performance and efficiency, this method has no fixed model parameters and relies on SPM statistical analysis. The inference time per frame is approximately 500ms (including 3D modeling and statistical correction), and the training cost is low—no large-scale model training is required, only data preprocessing. In terms of robustness, it has moderate anti-noise ability, is susceptible to ambient temperature, and requires a constant temperature environment of 23.5-24.5℃; its anti-occlusion ability is weak, as head movements or electrode occlusion can lead to the loss of thermal signals; and its environmental adaptability is weak, being only suitable for indoor controlled environments.

Applicable scenarios focus on medical diagnosis (emotional assessment for patients with autism spectrum disorder and Parkinson's disease) and the marketing field (research on consumers' emotional responses). The main limitations are strong dependence on equipment—specialized infrared thermal imaging cameras are required, resulting in high costs; high environmental sensitivity, as temperature and humidity can affect data accuracy; and semi-automated preprocessing, which requires manual alignment of 3D facial models.

3.1.2 Emotion-LLaMA (Multimodal Large Language Model for Emotion Recognition and Reasoning)

The data flow of its core technical framework is „construction of the MERR dataset (28,618 coarse-grained samples + 4,487 fine-grained samples) → multimodal feature extraction (HuBERT for audio processing, MAE/VideoMAE/EVA for visual processing, and LLaMA for text processing) → alignment of features into a unified space → instruction tuning". Key technologies include linear projection alignment of multimodal features, instruction tuning of the LLaMA model, and generalization enhancement through coarse-to-fine annotation.

The key innovations are integrating audio-visual-text multimodal emotional cues to cover both micro-expressions and contextual information; performing instruction tuning based on the MERR dataset to significantly improve the accuracy of emotion recognition and reasoning; and adopting a linear projection mechanism to align multimodal features with text tokens, enhancing cross-modal synergy.

In terms of performance and efficiency, the model parameters consist of 7B (LLaMA2-chat base model) + 34M trainable parameters (LoRA). The inference time per sample for multimodal input is approximately 200ms, and the training cost is high—it requires training 300,000 steps on 4×A100 GPUs, taking about 20 hours. In terms of robustness, it has strong anti-noise ability; relying on multimodal complementarity, noise in a single modality can be compensated by other modalities, achieving an F1 score of 0.8452 on the MER2024-NOISE dataset; its anti-occlusion ability is moderate—visual occlusion can be supplemented by audio/text, but it becomes ineffective in cases of severe occlusion; and its environmental adaptability is moderate, being suitable for both indoor and outdoor scenarios, though extreme lighting can affect the quality of visual features.

Applicable scenarios include human-computer interaction (emotional dialogue robots), educational tutoring (real-time analysis of students' emotional states), and mental health counseling (dynamic monitoring of visitors' emotions). The main limitations are the complex model structure, which requires substantial computing resources and makes deployment on edge devices difficult; high dependence on text, as emotion reasoning accuracy decreases in the absence of text; and the need for improved micro-expression recognition, as it is prone to confusion with complex expressions.

ISSN 2959-6157

3.1.3 EmotionNet-X (Optimized CNN Architecture for Robust Facial Emotion Analysis)

The data flow of its core technical framework is „image preprocessing (resizing to 48×48×1 + normalization) → feature extraction via 4 convolutional layers (with ReLU activation) → regularization via 7 Dropout layers → batch normalization → model training with the Adam optimizer". Key technologies include a lightweight CNN structure, multi-layer Dropout for overfitting suppression, batch normalization for training acceleration, and balancing parameter optimization and inference time.

The key innovations are reflected in its lightweight design, with only 19.9M parameters and an inference time of 18ms per image, making it suitable for resource-constrained devices; the combination of 4 convolutional layers + 7 Dropout layers + batch normalization to enhance robustness against noise and occlusion; and achieving an accuracy of 99.86% on the CK+ dataset, which is significantly better than that of pre-trained models such as VGG19 and ResNet50V2.

In terms of performance and efficiency, the model has 19.9M parameters, an inference time of 18ms per image, and low training cost—it requires training 500 epochs on a Tesla K80 GPU, taking about 8 hours. In terms of robustness, it has strong anti-noise ability; the combination of 7 Dropout layers and batch normalization enables it to maintain 62% accuracy on noisy FER2013 data; its anti-occlusion ability is moderate—it can recognize local occlusion but becomes ineffective in cases of full-face occlusion; and its environmental adaptability is moderate—it is robust to lighting changes but experiences a decrease in accuracy under extreme angles such as profile views.

Applicable scenarios include access control (emotion-assisted identity verification), health status monitoring (real-time early warning of abnormal facial emotions), and embedded security systems (deployment on resource-constrained devices). The main limitations are poor adaptability to non-facial images, as it only supports facial emotion analysis; limited coverage of emotion categories, with poor recognition performance for rare emotions (e.g., „contempt"); and dependence on fixed image size, being only compatible with 48×48×1 input.

3.1.4 Multimodal Emotion Analysis Based on Multi-Layer Feature Fusion and Multi-Task Learning

The data flow of its core technical framework is „feature extraction via the Unimodal Feature Extraction Network (UFEN) (multi-layer convolution + self-attention) → cross-modal interaction via the Multi-Task Fusion Network (MTFN) (multi-head attention + encoder-decoder) → parameter optimization via multi-task learning". Key technologies include UFEN, MTFN, cross-modal multi-head attention, and balancing cross-modal bias through multi-task loss.

The key innovations are combining multi-layer convolution and self-attention to extract fine-grained unimodal features, reducing interference from redundant information; using multi-head attention to achieve in-depth cross-modal interaction, fully leveraging the complementary advantages of different modalities; and optimizing parameters through multi-task learning to improve stability and accuracy in complex emotion analysis tasks.

In terms of performance and efficiency, the model has approximately 50M parameters (including UFEN and MTFN), an inference time of about 150ms per sample for multimodal input, and moderate training cost—it requires training 50 epochs on an RTX 3060Ti GPU, taking about 12 hours. In terms of robustness, it has strong anti-noise ability; the combination of multi-task learning and cross-modal interaction enables it to achieve a Mean Absolute Error (MAE) of 0.728 on the MOSI dataset; its anti-occlusion ability is moderate—missing parts of modalities can be compensated, but it becomes ineffective in cases of complete multimodal missing; and its environmental adaptability is moderate, being suitable for multiple scenarios, though performance decreases when modal synchronization is poor (e.g., audio-video asynchrony).

Applicable scenarios include public opinion analysis (judgment of emotional tendencies in multimodal public opinion), healthcare (multimodal emotional state assessment for patients), and education (multimodal analysis of students' learning emotions). The main limitations are the complex training process, which requires multimodal annotated data and long training time; high requirements for modal synchronization, as temporal/spatial bias affects fusion effectiveness; and poor interpretability, making it difficult to trace the multi-task loss optimization process.

3.2 Current Challenges and Future Prospects of Image-Based Emotion Analysis

3.2.1 Current Challenges

1. Emotion Expression Diversity & Environmental Interference: Human emotions vary by individual, culture, and scenario (e.g., mixed emotions like „anxious joy"); real-world factors (lighting, occlusion, head movement) further disrupt feature extraction. For example, pixel-based facial thermal imaging works in controlled environments but fails under extreme temperatures or heavy movement, while EmotionNet-X struggles with non-facial or occluded images.

2. Data & Annotation Issues: Datasets suffer from subjective annotation (inconsistent labels for the same image), high fine-grained annotation costs (e.g., MERR's 4,487

fine-grained samples require extensive labor), and severe class imbalance (rare emotions like „disgust“ are underrepresented), leading to biased models.

3. Multimodal Fusion Difficulty: Integrating heterogeneous image, audio, and text data is challenging—modalities have distinct features (spatial for images, temporal for audio) and may be asynchronous/incomplete. Methods like multi-layer feature fusion need complex encoder-decoder structures to align modalities, increasing complexity.

3.2.2 Future Prospects

1. Enhance Model Robustness & Generalization: Develop adaptive models that handle environmental interference (e.g., dynamic feature adjustment for occlusion/lighting) and generalize across populations (e.g., adapting to children's facial expressions) to bridge the gap between lab datasets (e.g., CK+) and real-world use.

2. Optimize Data & Efficient Models: Explore semi-supervised/unsupervised learning to reduce annotation reliance; design lightweight yet accurate models (balancing EmotionNet-X's speed and Emotion-LLaMA's accuracy) for real-time scenarios like intelligent driving.

3. Advance Multimodal Fusion: Focus on adaptive fusion (prioritizing reliable modalities dynamically) and cross-modal knowledge transfer (compensating for missing modalities) to fully leverage complementary image-audio-text information, improving analysis accuracy.

4. Conclusion

This paper comprehensively reviews key methods and applications of image-based emotion analysis, responding to the research goal of systematically organizing the field's progress proposed in the introduction. It covers core technologies including single-modal methods (facial thermal imaging with privacy protection, lightweight EmotionNet-X), improved ViT models (ViT-CoMer enhancing local features), and multi-modal fusion approaches (Emotion-LLaMA's cross-modal alignment, UFEN+MTFN's multi-task optimization).

These methods have achieved breakthroughs in accuracy (e.g., EmotionNet-X's 99.86% accuracy on CK+), efficiency (18ms inference time), and robustness (multi-modal anti-noise), but still face challenges like emotional diversity, data annotation issues, and multimodal fusion difficulties.

Current methods are insufficient in adapting to extreme environments; future research should enhance model generalization, explore semi-supervised data solutions, and advance adaptive multimodal fusion technologies to promote practical application.

Authors Contribution

All the authors contributed equally, and their names were listed in alphabetical order.

References

[1] Schuller BW, Schuller DM. Audiovisual affect assessment and autonomous automobiles: Applications. *arXiv preprint arXiv:2203.07482*. 2022.

[2] Pilli S, Patwardhan M, Pedanekar N, Karande S. Predicting sentiments in image advertisements using semantic relations among sentiment labels. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2020. p. 408–9.

[3] Budu T, Sato W, Shimokawa K, Hsu CT, Kochiyama T. Development of pixel-based facial thermal image analysis for emotion sensing. *Computers in Human Behavior Reports*. 2025;100761.

[4] Sato W, Kochiyama T, Yoshikawa S. Physiological correlates of subjective emotional valence and arousal dynamics while viewing films. *Biological Psychology*. 2020;157:107974.

[5] Cheng Z, Cheng ZQ, He JY, Wang K, Lin Y, Lian Z, et al. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. *Advances in Neural Information Processing Systems*. 2024;37:110805–53.

[6] Abbas SMA, Abbas Q, Naqi SM. EmotionNet-X: An optimized CNN architecture for robust facial emotion analysis. *IEEE Access*. 2025.

[7] Cai Y, Li X, Zhang Y, Li J, Zhu F, Rao L. Multimodal sentiment analysis based on multi-layer feature fusion and multi-task learning. *Scientific Reports*. 2025;15(1):2126.

[8] Sagaidachnyi AA, Fomin AV, Usanov DA, Skripal AV. Thermography-based blood flow imaging in human skin of the hands and feet: A spectral filtering approach. *Physiological Measurement*. 2017;38(2):272.

[9] Xia C, Wang X, Lv F, Hao X, Shi Y. Vit-comer: Vision transformer with convolutional multi-scale feature interaction for dense predictions. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024. p. 5493–502.

[10] Haruna Y, Qin S, Chukkol AHA, Yusuf AA, Bello I, Lawan A. Exploring the synergies of hybrid convolutional neural network and Vision Transformer architectures for computer vision: A survey. *Engineering Applications of Artificial Intelligence*. 2025;144:110057.

[11] Nadeem M, Sohail SS, Javed L, Anwer F, Saudagar AKJ, Muhammad K. Vision-enabled large language and deep learning models for image-based emotion recognition. *Cognitive Computation*. 2024;16(5):2566–79.