ISSN 2616-5775 Vol. 9, Issue 1: 1-11, DOI: 10.25236/AJCIS.2026.090101

A Review of Deep Learning in Micro-expression Recognition Research

Mincong Zhou[1], Wei Wang[2,*], Yixing Li[1]

[1]School of Health Science and Engineering, University of Science and Technology, Shanghai, China
[2]PLA Naval Medical Center, Naval Medical University, Shanghai, China
**Corresponding author: wwang_fd@fudan.edu.cn*

Abstract: *Micro-expressions (ME) are involuntary, fleeting facial cues that reveal hidden emotions in high-stakes situations. They provide valuable insights into an individual's true psychological state and have a wide range of applications in psychology, law enforcement, and human-computer interaction. Traditional ME recognition relies on hand-crafted features, but recent advances in deep learning have made end-to-end recognition possible, greatly accelerating research progress. This paper reviews deep learning-based micro-expression recognition (MER), covering dataset construction, pre-processing, feature enhancement, and the evolution of network architecture. It also compares single-stream, multi-stream, and multimodal fusion models, summarizes loss strategies, and discusses current challenges and future research trends. This review aims to provide a systematic perspective to promote the development and practical reliability of MER.*

Keywords: *Micro-expression Recognition, Feature Detection, Feature Extraction, Micro-expression dataset, Deep Learning*

1. Introduction

Micro-expressions (MEs) are fleeting, involuntary facial movements that occur when individuals attempt to conceal genuine emotions under high-pressure or high-stakes situations. The concept of micro-expressions (MEs) was first introduced by Haggard and Isaacs (1966)[1], who identified them in psychotherapy video recordings as manifestations of repressed emotions and initially referred to them as “rapid expressions”. Later, Ekman and Friesen (1969), while analyzing video footage of a psychiatric patient with suicidal tendencies, observed extremely brief facial movements that reflected concealed emotions. They formally defined these fleeting facial behaviors as “micro-expressions”. They provide valuable cues to a person's true psychological state and have broad applications in psychology, national security, clinical diagnosis, and human-computer interaction. Due to their extremely brief duration (typically 1/25-1/5s), low intensity, and localized manifestation, MEs are difficult to observe with the naked eye, making their automatic recognition an active research topic in computer vision.

With advancements in deep learning, micro-expression recognition (MER) has evolved from traditional manual feature methods such as LBP-TOP [2] and optical flow to end-to-end frameworks using convolutional, looping, and neural networks. These methods realize the joint extraction and multimodal fusion of spatiotemporal features, which greatly improves the recognition accuracy. However, challenges remain, including limited dataset size, between-subject variations, and generalization across documentation conditions. Therefore, this study aims to provide a comprehensive review of deep learning-based MER research.

2. Dataset

During the collection of micro-expression databases, participants are typically instructed to maintain a neutral facial expression while viewing emotion-inducing videos, thereby eliciting spontaneous micro-expression. Existing micro-expression databases primarily fall into two categories: early databases and laboratory-based spontaneous databases. The former, such as the York-DDT, Polikovsky and USF-HD databases, predominantly rely on participants mimicking or rapidly performing expressions. The latter capture spontaneous micro-expression through emotional induction, such as SMIC, the CASME series, SAMM, MMEW, and the latest CAS(ME)^3 . Table 1 presents the distinctions among these datasets.

2.1. Early databases

York-DDT (2009)[3]: This was the first database related to lie detection, involving 20 participants. Researchers played two video clips, one designed to elicit an emotional response and the other lacking emotional induction. Participants were instructed to describe the first clip as “deceptive” and the second as “truthful”. This study supported the theory of non-verbal leakage. However, the database contains substantial extraneous head and facial movements, resulting in high noise levels that render it unsuitable for precise micro-expression recognition research.

Polikovsky Database (2009)[3]: Captured using a Point Grey Grasshopper camera at 480×640 resolution and 200 fps (RAW8 mode). This database is not publicly available, and limited information exists regarding it, such as the undisclosed total sample size. During collection, ten participants were instructed to “imitate” micro-expression. This performative data lacks genuine ecological validity, constituting its primary shortcoming.

USF-HD (2011)[4]: Contains 181 macro-expression and 100 micro-expression samples, captured using JVC-HD100 or Panasonic AG-HMC40 cameras at 720×1280 resolution and approximately 29.7 fps. Although the dataset is slightly larger than York-DDT and Polikovsky, it still contains a significant amount of imitation and lacks authenticity.

2.2. Common databases

SMIC (2013)[5]: Released by the University of Oulu, Finland, comprising 20 subjects. Data was synchronously captured using three distinct cameras: ① High-speed camera (HS, 100 fps, 640×480, 164 samples); ② Standard visible-light camera (VIS, 25 fps, 71 samples); ③ Near-infrared camera (NIR, 25 fps, 71 samples). All data were captured as full-face frontal views, categorized into positive, negative, and surprise emotional states. The database's strength lies in its multimodal acquisition, though it features a limited range of emotional categories.

CASME (2013)[6]: Developed by the Institute of Psychology, Chinese Academy of Sciences, involving 35 participants. Of these, 19 produced valid micro-expression, yielding a total of 195 samples. Data acquisition comprises two classes: Class A (BenQ M31, 1280×720, 60 fps, 100 samples) and Class B (Point Grey GRAS-03K2C, 640×480, 60 fps, 95 samples). Emotion categories comprised eight types: Pleasure, Contempt, Disgust, Fear, Sadness, Surprise, Suppression, and Tension.

CASME II (2014)[7]: An upgraded version of CASME, containing 247 samples from 26 subjects. Captured using Point Grey GRAS-03K2C at 200 fps, 640×480 resolution, with facial region cropped to 280×340. Emotion labels comprise 5 categories: disgust, pleasure, surprise, suppression, and other, alongside onset, apex, and offset frame annotations plus action unit (AU) labelling. This database, characterized by its high frame rate and meticulous annotation, has become the most frequently used benchmark for MER.

CAS(ME)^2(2017)[7]: Constructed from 22 subjects using a Logitech C920 camera (30 fps, 640×480). The dataset comprises 300 macro-expressions and 57 micro-expressions categorized as positive, negative, surprise, and other. Its distinctive feature is the inclusion of “long videos” making it suitable for combined research on micro-expression spotting and recognition.

SAMM (2016)[8]: Developed by a UK team, comprising 159 samples from 32 participants. Recorded using a Basler Ace acA2000-340km high-speed camera (200 fps, 2040×1088) with an LED array to eliminate AC flicker. Seven emotional categories: anger, sadness, contempt, disgust, fear, joy, and surprise. The database's strengths lie in its coverage of 13 ethnicities, strong racial diversity, and annotation by FACS experts.

MEVIEW (2019)[9]: Collected from YouTube videos (e.g., poker tournaments, television interviews), comprising 31 video segments from 16 individuals. This database represents in-the-wild scenarios, holding significant practical value. However, due to its small sample size and high video noise, it is currently primarily used for exploratory research.

MMEW (2020)[10]: Collected by a Chinese team, comprising 300 micro-expressions and 900 macro-expressions from over 40 participants. Resolution: 1920×1080, frame rate: 90 fps. Micro-expression and macro-expressions originate from the same participants, facilitating cross-modal research.

CAS(ME)^3 (2023)[11]: Recently released by the Institute of Psychology, Chinese Academy of Sciences, this is currently the largest multimodal micro-expression database. It comprises 216 participants, approximately 943 micro-expressions, and 3,143 macro-expressions, spanning over 80 hours of footage. Captured modalities include RGB images, depth maps, audio (48 kHz), and physiological signals (EDA, ECG, RSP, PPG), providing invaluable resources for cross-modal research.

Table 1: Summary of micro-expression datasets.

Dataset	Sub	Resolution	FR	Face size	Expression
CASME	35	640×480 1280×720	60	150×90	Hap(5), Dis(88), Fea(2), Sad(6), Sur(20), Rep(40), Ten(28), Con(3)
CASME II	26	640×480	200	250×340	Hap(33), Dis(60), Sur(25), Rep(27), Oth(102)
CAS(ME)^2	22	640×480	30	-	Hap(51), Sur(43), Neg(70), Oth(19)
CAS(ME)^3	247	1280×720	30	-	Hap(992), Dis(2528), Fea(892), Ang(619), Sad(635), Sur(1208), Sur(1208), Oth(251)
SMIC	16	640×480	100	190×230	Pos(51), Neg(70), Sur(43)
SAMM	32	2040×1088	200	400×400	Hap(24), Sur(13), Sur(20), Dis(8), Fea(7), Sad(3), Oth(84)
MMVIEW	16	720×1280	30	-	Hap(6), Ang(2), Sur(9), Dis(1), Fea(3), Unc(13),Con(102)
MMEW	36	1920×1280	90	400×400	Hap(36),Ang(8), Sur(80), Dis(72), Fea(16),Sad(13), Oth(102)

Sub = subjects, Hap = happiness, Ang = Anger, Sur = Surprise, Fea = Fear, Sad = Sadness, Dis = Disgust, Pos = Positive, Neg = Negative, Rep = Repression, Con = Contempt, Ten = Tense, Unc = Unclear, Oth = others.

3. Pre-processing and Feature Enhancement

Pre-processing serves as a crucial stage that bridges raw data acquisition and effective feature learning. Through procedures such as face detection, alignment, motion amplification, and illumination normalization, it aims to eliminate redundant variations and highlight subtle facial movements essential to micro-expression recognition. In addition, certain feature enhancement techniques further amplify discriminative motion cues and improve the robustness of subsequent learning stages. Together, these processes lay a solid foundation for the following feature extraction and representation learning stages.

3.1. Face Detection and Alignment

In the facial preprocessing stage, face detection and face alignment are two critical steps ensuring the successful execution of micro-expression recognition. The former primarily addresses the question of “where is the face?” by swiftly and accurately locating facial regions within images or videos to eliminate background interference. The latter tackles “how to standardize the face?” through key point localization and geometric normalization. This reduces the interference of variations in pose, lighting, and expression on subsequent feature extraction. The process of the micro-expression recognition is shown in Figure 1.

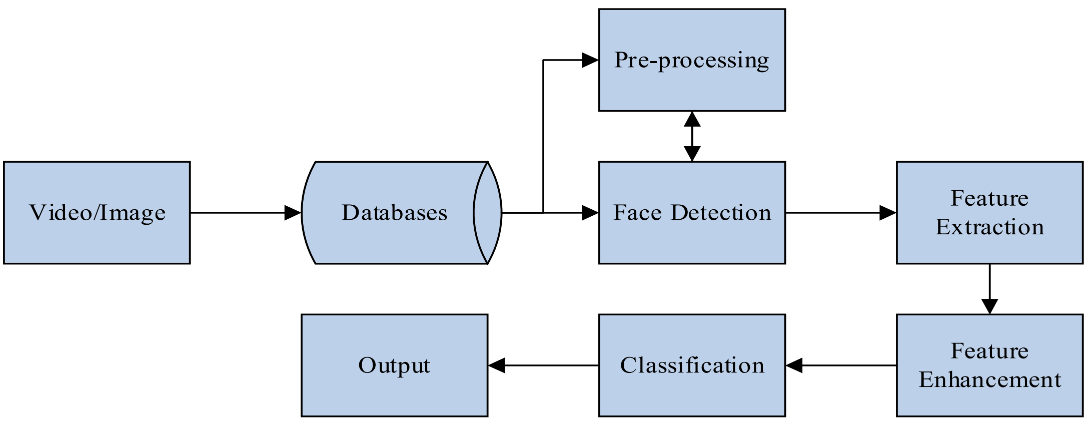


Figure 1: The whole process of micro-expression recognition

3.1.1. Face Detection Technology

In the field of face detection, research methodologies have evolved from traditional approaches to deep learning. The early Viola-Jones algorithm (Viola & Jones, 2001)[12], based on Hair features and cascaded AdaBoost classifiers, achieved real-time face detection but lacked robustness to pose and illumination variations. Subsequently, researchers introduced detection methods grounded in statistical modelling, such as skin tone modelling and geometric feature detection. However, these approaches proved constrained by complex backgrounds and have consequently seen reduced application in micro-expression recognition (MER). In recent years, deep learning approaches have increasingly become mainstream. Among these, MTCNN (Multi-task Cascaded CNN) (Zhang et al., 2016)[13] employs a cascaded multi-task network to achieve face detection and coarse alignment, finding widespread adoption in preprocessing for databases such as CASME II and SAMM. Additionally, detectors based on deep convolutional networks, such as RetinaFace[14] and Face++[15], have demonstrated high robustness. Overall, deep learning approaches have become the mainstream choice for facial detection in MER due to their advantages in complex backgrounds, pose variations, and occlusion conditions.

3.1.2. Face Alignment Methods

Following face detection, the detected facial regions require alignment to mitigate the impact of pose and expression variations. Early approaches primarily relied on statistical modelling. For instance, the Active Shape Model (ASM)[16] constrained key point locations by constructing shape models, while the Active Appearance Model (AAM)[17] further integrated texture information to model the overall facial appearance. These approaches were commonly employed in preprocessing for early databases like CASME and SMIC, yet exhibited sensitivity to lighting and occlusion. To enhance alignment accuracy, researchers proposed lightweight regression-based methods such as Constrained Local Model (CLM) [18]. In recent years, deep learning approaches have gained prominence. TCDCN (Tasks-Constrained Deep CNN)[19] employs multi-task learning to enhance landmark localization, while DAN (Deep Alignment Network)[20] progressively refines landmarks through cascaded regression, while HRNet[21] maintains high-resolution features for refined alignment. These approaches demonstrate superior robustness and accuracy in experiments on databases such as CASME II and SAMM, significantly enhancing the reliability of region of interest (ROI) segmentation and subsequent feature modelling.

3.2. Motion Magnification

A key challenge with micro-expressions lies in their minute amplitude, rendering subtle facial muscle movements difficult to observe. Consequently, motion magnification techniques have become a crucial step in enhancing micro-expression detectability. Currently, primary motion amplification methods include Eulerian Video Magnification (EVM), Lagrangian Motion Magnification and deep learning-driven amplification approaches.

3.2.1. EVM

Eulerian Video Magnification (EVM)[21], originally developed by the Massachusetts Institute of Technology (MIT), operates by decomposing video sequences into spatial pyramids and applying temporal bandpass filtering to pixel intensity variations.

Within the context of micro-expression recognition, EVM serves two primary functions. First, it amplifies subtle local facial muscle movements. Second, it strengthens fine-grained dynamic features, improving the sensitivity of feature extraction techniques to minor motion cues. Nevertheless, EVM entails considerable computational complexity and remains sensitive to noise, illumination fluctuations, and unintended motion. Excessive magnification factors may further introduce artifacts or false motion, underscoring the necessity for careful parameter optimization and integration with complementary processing techniques.

3.2.2. LMM

Lagrangian Motion Magnification (LMM) differs from Eulerian methods in that it directly amplifies subtle facial expressions based on the motion trajectories of pixels or feature points. Its fundamental principle involves tracking key points or local regions within facial sequences, calculating displacement changes over time, and applying magnification at the motion trajectory level. Flotho et al. [22] employed key techniques including forward deformation and optical flow fine-tuning in their local Lagrangian motion magnification method, significantly improving the accuracy and reliability of micro-expression amplification Notably, Global Lagrangian Motion Magnification (GLMM)[23] achieves more stable overall facial dynamic enhancement by applying consistent global tracking and magnification across the entire face. For micro-expression recognition, LMM/GLMM effectively amplifies subtle displacements in local regions such as eyebrows, corners of the mouth, or eyes. However, this method relies heavily on the accuracy of motion tracking. Errors can arise when head movement or occlusion occurs, leading to diminished amplification effects.

3.3. Temporal Normalization

The duration of micro-expression varies considerably, typically ranging from tens to hundreds of milliseconds. Directly feeding these raw sequences into feature extraction or classification models may result in redundant information within longer sequences and insufficient dynamic information in shorter ones, thereby compromising recognition performance. Consequently, temporal normalization emerges as a critical step in micro-expression preprocessing. The most classical approach is the Temporal Interpolation Model (TIM)[24], which employs graph embedding techniques to interpolate sequences on a low-dimensional manifold. However, TIM is prone to introducing redundancy or spurious motion when over-interpolating.

4. Method

The introduction of deep learning has brought significant advances to micro-expression recognition (MER). Compared to traditional handcrafted feature descriptors (such as LBP-TOP and HOG), deep models can automatically learn high-level semantic features through an end-to-end approach, better capturing subtle and transient emotional shifts. In recent years, researchers have advanced deep learning studies in MER on two fronts: firstly, at the module level, focusing on the feature learning capabilities of different network components; secondly, at the model architecture level, exploring how to organize and integrate diverse modules within the overall system to form higher-level representations

4.1. Network Modules

Feature extraction, as the core component of the micro-expression recognition process, directly determines the model's ability to perceive, model, and discriminate facial expression signals. In micro-expression recognition tasks, common network modules such as CNN, RNN, 3DCNN, and Transformer each possess distinct advantages and applicable scenarios.

4.1.1. CNN

Convolutional neural networks[25] were the earliest deep learning modules adopted in MER research. Figure 2 illustrates the architecture of the Convolutional Neural Network (CNN). Early studies utilized apex frames or key frames as input, extracting spatial texture features through two-dimensional convolutions, and combined these with fully connected layers or traditional classifiers to accomplish recognition. The primary breakthrough of this approach lay in replacing traditional manual feature extraction with automatic learning, significantly enhancing feature discriminability. To overcome overfitting issues arising from data scarcity, researchers commonly employ transfer learning strategies, adapting pre-trained networks such as VGG or ResNet to MER tasks. Some studies further incorporate difference images, optical flow maps, or dynamic images to endow CNNs with temporal sensitivity within static structures. While the introduction of CNN modules demonstrates the efficacy of deep features in micro-expression recognition, their capacity for dynamic modelling remains limited, capturing only spatial-level facial changes.

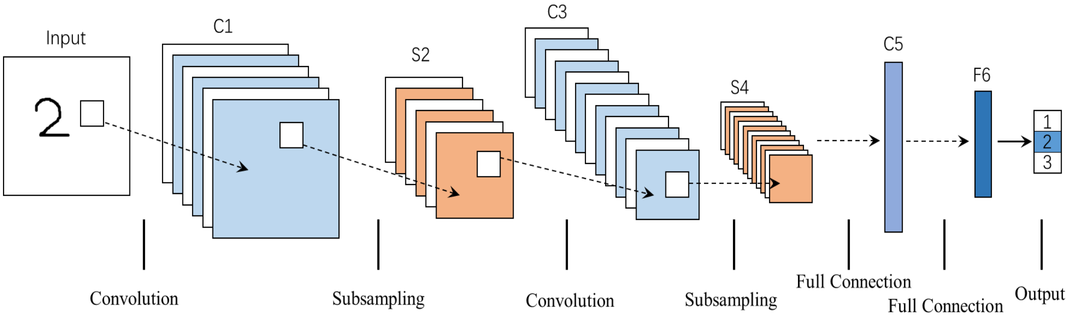
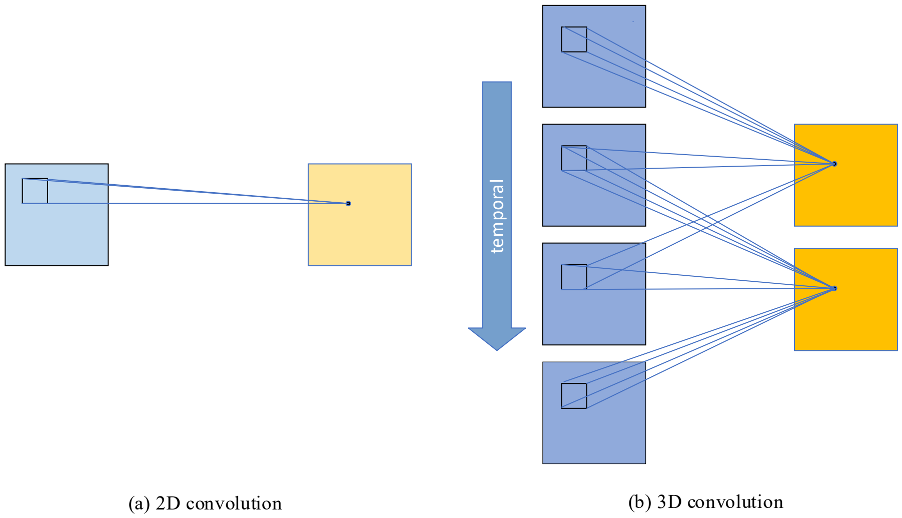


Figure 2: CNN model diagram

4.1.2. 3D-CNN

As research progressed, three-dimensional convolutional neural networks(3DCNN)[26] were proposed to simultaneously model spatial and temporal features. Figure 3 compares and contrasts 3D convolution with 2D convolution. Three-dimensional convolutions enable learning of local motion patterns across spatiotemporal domains, thereby more accurately capturing subtle variations in micro-expressions across their onset-apex-offset phases. On high-frame-rate datasets such as CASME II and SAMM, 3D-CNNs demonstrate markedly superior performance to 2D CNNs, particularly in recognizing low-intensity and suppressed emotions. To enhance model robustness, several approaches combine temporal interpolation with motion amplification techniques to mitigate the impact of sparse samples during training. Nevertheless, the high computational cost and substantial parameter count of 3D convolutional networks remain significant bottlenecks.



(a) 2D convolution

(b) 3D convolution

Figure 3: 2D convolution and 3D convolution

4.1.3. RNN

Recurrent neural networks(RNNs)[27] and their variants are primarily employed for time series modelling in MER. Figure 4 illustrates the architecture of the Recurrent Neural Network (RNN). A typical architecture adopts the CNN-RNN framework: CNNs extract spatial features from individual frames, while RNNs (typically LSTM[28] or GRUs) capture temporal dependencies. This integration enables the model to learn contextual information about facial expression changes from consecutive frames, demonstrating particular efficacy in recognizing subtle movements during dynamic transitional phases. To enhance temporal awareness, some studies employ bidirectional LSTMs to capture both forward and backward temporal dependencies. Others introduce temporal attention modules, assigning higher weights to key frames near the apex. While such architectures offer advantages in dynamic modelling, their serial computation mechanism results in inefficient training and susceptibility to gradient vanishing issues with lengthy sequences.

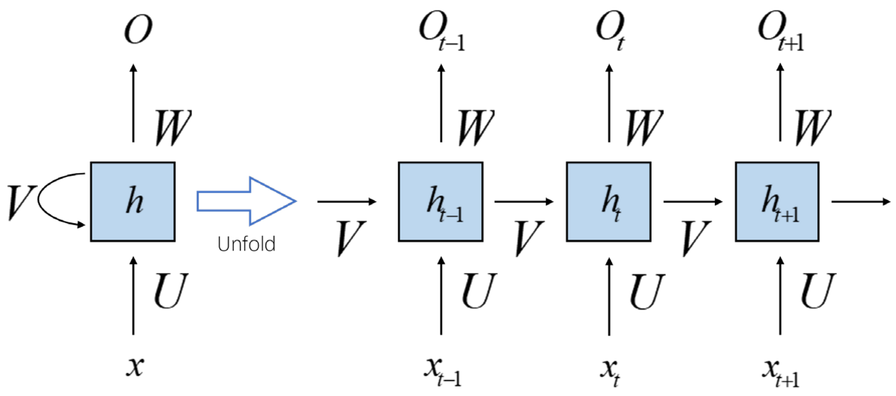


Figure 4: RNN Model diagram

4.1.4. Transformer

In recent years, the introduction of Transformer modules has brought novel approaches to global modelling in MER. Through self-attention mechanisms, Transformers[29] can capture global temporal dependencies, overcoming the limitations of convolutional and recurrent structures in local modelling. Figure 5 illustrates the architecture of the Transformer. In specific implementations, researchers typically treat frame-by-frame CNN features or spatiotemporal patches as input tokens, learning dependencies between different frames or regions via multi-head attention layers. Some studies have integrated spatial and temporal attention mechanisms, enabling models to focus both on key facial regions (such as muscle groups corresponding to AUs) and critical moments of expression change. Experimental results demonstrate the Transformer's exceptional generalize capability across database tasks (e.g., CASME II →SAMM). By integrating self-supervised learning, contrastive learning, and meta-learning strategies, the Transformer further enhances model robustness under small-sample conditions. Concurrently, its attention visual properties provide interpretable support for model or decision-making.

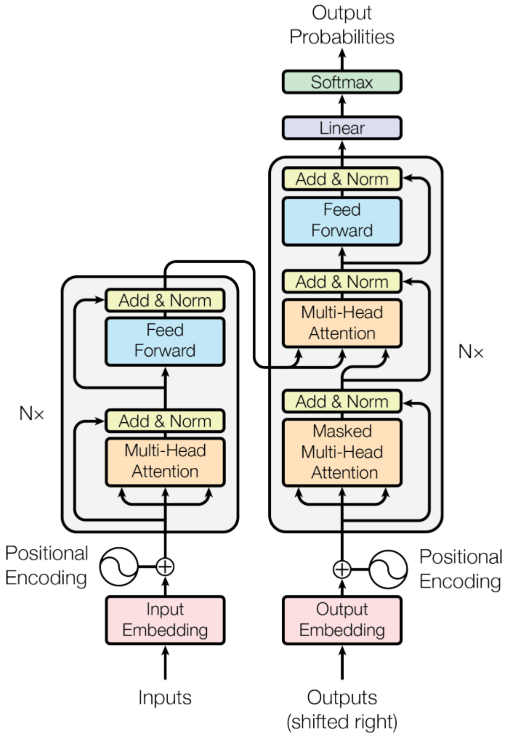


Figure 5: Transformer Model diagram.

4.2. Model Architectures

The practical challenges confronting micro-expression recognition tasks include sparse sample quantities, extremely subtle expression magnitudes, and inadequate model generalization capabilities across different individuals. Consequently, model architecture design must not only focus on enhancing expressive capacity but also consider computational efficiency, parameter scale, and robustness. Recent research indicates that rational network structure design can effectively improve model performance under limited data conditions.

4.2.1. Single-Stream Architectures

The single-stream architecture represents the most fundamental form of deep learning framework within MER, where models perform end-to-end modelling from input to classification through a single pathway. Early CNN[25], 3D-CNNs[26], and CNN-RNNs all belong to this category. Its advantages lie in structural simplicity and controllable parameters, enabling stable training on limited datasets. As model complexity increased, researchers introduced hybrid convolutional blocks and multi-layer attention mechanisms into single-stream architectures to enhance dynamic modelling capabilities within constrained computational resources.

Chen et al. (2020)[30] proposed the Spatiotemporal Convolutional Network with Convolutional Block

Attention Module (CBAM-STCNN), which is an exemplary optimization of a single-stream model. This model embeds channel and spatial attention mechanisms into a single 3D convolutional network, enabling it to adaptively focus on key spatiotemporal features of micro-expressions without relying on a multi-stream design, achieving efficient and accurate recognition.

Gajjala et al. (2021)[31] proposed MERANet (3D Residual Attention Network), which also belongs to an advanced single-stream architecture. This network deeply integrates 3D residual learning with a hybrid attention mechanism, ensuring gradient flow and feature reuse through residual connections while refining features using attention, significantly enhancing the ability to capture subtle motion information.

Liu et al. (2022)[32] designed the SQU-C3D model, which is a single-stream network. Its core innovation lies in the keyframe sampling strategy: first, the lightweight SqueezeNet is used to locate the vertex frame that best represents the strength of micro-expression, and then the sequence consisting of three keyframes, the start, vertex frame, and end frame, is input into the C3D network to efficiently learn the spatiotemporal characteristics of micro-expression.

Wang et al.(2023)[33] developed HTNet (Hierarchical Transformer Network), a single-stream architecture that effectively captures expressions from local subtle features to global co-motion by dividing faces into four key regions and performing hierarchical attention and block aggregation within the network.

4.2.2. Multi-Stream Architectures

As research progressed, scholars increasingly recognized that a single information source struggles to comprehensively characterize micro-expressions. Multi-stream architectures consequently emerged as a novel research direction. The most typical two-stream network comprises an appearance stream and a motion stream: the former processes RGB images to capture texture features, while the latter processes optical flow maps or dynamic images to model motion characteristics. Subsequently, multi-stream networks have progressively expanded into tri-stream or multi-modal configurations, integrating depth images, thermal infrared imagery, audio signals, or physiological data to achieve more comprehensive emotional analysis.

Khor et al. (2019)[34] proposed the Dual-Stream Shallow Network (DSSN), which adopts a typical dual-stream heterogeneous architecture. This network constructs a pair of lightweight convolutional neural networks to process different motion feature inputs (such as optical flow magnitude and optical strain magnitude) separately, and performs feature fusion in the later stages of the network, effectively combining the advantages of different motion representations and improving recognition performance while maintaining parameter lightweight.

Liong et al. (2019)[35] proposed the Shallow Triple Stream Three-dimensional CNN (STSTNet), extending the multi-stream architecture into three-dimensional space. This model uses three parallel shallow 3D CNNs to process the original sequence, horizontal optical flow, and vertical optical flow separately, extracting spatiotemporal features from different dimensions and finally fusing them, decomposing and capturing facial movement patterns in a more refined manner.

Li et al. (2022)[36] proposed the deep local-global network (DLHN). The network consists of two parallel subnetworks: HCRNN focuses on extracting local, fine spatiotemporal features from multiple regions of interest (ROIs) on the face, and gradually extracts local features at different scales through hierarchical convolutional structure. RPRNN uses robust principal component analysis (RPCA) technology to extract global and sparse micro-expression motion information.

Zhu et al. (2024)[37] proposed a three-stream temporal offset attention network (SKD-TSTSAN) based on self-knowledge distillation, which includes static spatial flow, local space, and dynamic temporal flow, and further optimized the feature learning ability through self-knowledge distillation technology.

4.2.3. Multimodal Fusion Frameworks

Most existing micro-expression recognition (MER) research has focused on learning features sensitive to facial expressions. However, in the real world, micro-expressions are often intertwined with various factors such as subject identity and action units (AUs). Approaches tailored solely to a single MER task fail to fully use facial information. The multimodal fusion method is conceptually broader than the multi-stream model, emphasizing the integration of data modalities with greater intrinsic differences (such as visual features and physical motion features, data distribution in different databases) to solve more complex problems such as generalization.

Zhou et al. (2019)[38] proposed the Dual-Inception Network, focusing on addressing the challenge of cross-database recognition, which can be seen as a fusion aimed at data distribution modalities. This network utilizes multi-scale feature extraction and domain adaptation techniques to reduce distribution differences between different databases, improving the model's generalization ability on unseen data.

The hierarchical spatiotemporal attention model (HSTA) proposed by Hao et al. (2024)[39] represents a recent advancement in multimodal fusion technology. The model adopts a hierarchical architecture design and includes two core innovation modules: the single-modal spatiotemporal attention (USTA) module performs internal feature extraction and temporal relationship modeling for video frames and special frames (such as vertex frames) respectively; The Cross-Modal Spatiotemporal Attention (CSTA) module realizes deep intelligent information fusion through the cross-attention mechanism, so that the features of one modality (such as special frames) can actively query and enhance the features of another modality (such as video frames).

4.3. Loss function

Within deep learning frameworks, loss functions not only determine the convergence direction of models but also directly influence the distribution patterns within feature spaces. Micro-expression recognition is characterized by substantial intra-class variation, minimal inter-class distinction, and uneven data distribution. Consequently, designing appropriate loss functions is crucial for enhancing recognition accuracy.

Presently, most MER models still employ Softmax cross-entropy loss[40] as their primary optimization objective, enabling end-to-end classification under known category conditions. However, reliance solely on Softmax often fails to guarantee sufficient feature discriminative power and generalization capability. Consequently, researchers have proposed various metric learning and marginal constraint loss approaches to address this limitation.

Centre loss[41] represents another prevalent enhancement approach. By penalizing the distance between deep features and their corresponding class centers, center loss constrains intra-class feature aggregation, thereby enhancing discriminative power. Furthermore, addressing the common long-tail distribution issue in micro-expression datasets, researchers introduced Focal Loss. This method effectively mitigates bias from class imbalance by increasing the attention weight on hard-to-classify samples. The MER-GCN model further builds upon Focal Loss by designing an adaptive factor to dynamically adjust the weight of positive and negative samples within training batches, thereby achieving more balanced gradient updates.

5. Future and Outlook

Despite deep learning's significant advancement in micro-expression recognition (MER), limitations persist in data scarcity, cross-database generalization, interpretability, and practical application. Future research should deepen efforts in constructing high-quality databases and weak supervision learning, mitigating data scarcity through multimodal acquisition, self-supervised learning, and generative models. At the model level, cross-domain generalization and adaptability should be enhanced, leveraging domain adaptation and meta-learning to achieve robust feature representations. This approach simultaneously enhances model interpretability and lightweight performance through action unit (AU) analysis, attention visualization, and knowledge distillation, thereby achieving efficient and transparent decision-making. Furthermore, it expands the application boundaries of Multimodal Emotion Recognition (MER) via multimodal and multi-task fusion, enhancing its practical value and societal significance in psychological assessment, human-computer interaction, and affective computing.

References

[1] Haggard E A, Isaacs K S. Micromomentary facial expressions as indicators of ego mechanisms in psychotherapy[M]. Methods of Research in Psychotherapy. Springer, 1966: 154-165.

[2] Zhao G, Pietikainen M. Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(6): 915-928.

[3] Polikovsky S, Kameda Y, Ohta Y. Facial micro-expressions recognition using high speed camera and 3D-gradient descriptor[C].3rd International Conference on Imaging for Crime Detection and Prevention (ICDP 2009). 2009: 1-6.

[4] Shreve M, Godavarthy S, Goldgof D, et al. Macro- and micro-expression spotting in long videos using spatio-temporal strain[C].2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG). 2011: 51-56.

[5] Li X, Pfister T, Huang X, et al. A Spontaneous Micro-expression Database: Inducement, collection and baseline[C].2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG). 2013: 1-6.

[6] Yan W J, Wu Q, Liu Y J, et al. CASME database: A dataset of spontaneous micro-expressions collected from neutralized faces[C].2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG). 2013: 1-7.

[7] Yan W J, Li X, Wang S J, et al. CASME II: An Improved Spontaneous Micro-Expression Database and the Baseline Evaluation[J]. PLoS ONE, 2014, 9(1): e86041.

[8] Davison A K, Lansley C, Costen N, et al. SAMM: A Spontaneous Micro-Facial Movement Dataset[J]. IEEE Transactions on Affective Computing, 2018, 9(1): 116-129.

[9] Husak P, Cech J, Matas J. Spotting Facial Micro-Expressions " In the Wild "[C]. Proceedings of the Computer Vision Winter Workshop. Retz, Austria; 2017: 1-9.

[10] Ben X, Ren Y, Zhang J, et al. Video-Based Facial Micro-Expression Analysis: A Survey of Datasets, Features and Algorithms[J]. IEEE transactions on pattern analysis and machine intelligence, 2022, 44(9): 5826-5846.

[11] Li J, Dong Z, Lu S, et al. CAS(ME)3: A Third Generation Facial Spontaneous Micro-Expression Database With Depth Information and High Ecological Validity[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(3): 2782-2800.

[12] Viola P, Jones M. Robust real-time face detection[C].Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001: Vol. 2. 2001: 747.

[13] Yin X, Liu X. Multi-Task Convolutional Neural Network for Pose-Invariant Face Recognition[J]. IEEE Transactions on Image Processing, 2018, 27(2): 964-975.

[14] Deng J, Guo J, Zhou Y, et al. RetinaFace: single-stage dense face localisation in the wild[A]. arXiv, 2019.

[15] Zhou E, Fan H, Cao Z, et al. Extensive facial landmark localization with coarse-to-fine convolutional network cascade[C].Proceedings of the IEEE International Conference on Computer Vision Workshops. 2013: 386-391.

[16] Cootes T F, Taylor C J, Cooper D H, et al. Active Shape Models-Their Training and Application[J]. Computer Vision and Image Understanding, 1995, 61(1): 38-59.

[17] Cootes T F, Edwards G J, Taylor C J. Active appearance models[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001, 23(6): 681-685.

[18] Cristinacce D, Cootes T F. Feature Detection and Tracking with Constrained Local Models[C]. Procedings of the British Machine Vision Conference 2006. Edinburgh: British Machine Vision Association, 2006: 95.1-95.10.

[19] Zhang Z, Luo P, Loy C C, et al. Facial Landmark Detection by Deep Multi-task Learning[C].Fleet D, Pajdla T, Schiele B, et al. Computer Vision – ECCV 2014. Cham: Springer International Publishing, 2014: 94-108.

[20] Kowalski M, Naruniec J, Trzcinski T. Deep Alignment Network: A convolutional neural network for robust face alignment[A]. arXiv, 2017.

[21] Wang J, Sun K, Cheng T, et al. Deep High-Resolution Representation Learning for Visual Recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(10): 3349-3364.

[22] Flotho P, Heiss C, Steidl G, et al. Lagrangian motion magnification with double sparse optical flow decomposition[J]. Frontiers in Applied Mathematics and Statistics, 2023, 9: 1164491.

[23] Le Ngo A C, Johnston A, Phan R C W, et al. Micro-Expression Motion Magnification: Global Lagrangian vs. Local Eulerian Approaches[C].2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). 2018: 650-656.

[24] Zhou Z, Zhao G, Pietikäinen M. Towards a practical lipreading system[C].CVPR 2011. 2011: 137-144.

[25] Lecun Y, Bottou L, Bengio Y, et al. Gradient-Based Learning Applied to Document Recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[26] Tran D, Bourdev L, Fergus R, et al. Learning Spatiotemporal Features with 3D Convolutional Networks[A]. arXiv, 2015.

[27] Lipton Z C, Berkowitz J, Elkan C. A Critical Review of Recurrent Neural Networks for Sequence Learning[A]. arXiv, 2015.

[28] Hochreiter S, Schmidhuber J. Long Short-Term Memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[29] Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need[A]. arXiv, 2023.

[30] Chen B, Zhang Z, Liu N, et al. Spatiotemporal Convolutional Neural Network with Convolutional Block Attention Module for Micro-Expression Recognition[J]. Information, 2020, 11(8): 380.

[31] Gajjala V R, Reddy S P T, Mukherjee S, et al. MERANet: facial micro-expression recognition using 3D residual attention network[C].Proceedings of the Twelfth Indian Conference on Computer Vision, Graphics and Image Processing. New York, NY, USA: Association for Computing Machinery, 2021: 1-10.

[32] Liu S, Ren Y, Li L, et al. Micro-expression recognition based on SqueezeNet and C3D[J]. Multimedia Systems, 2022, 28(6): 2227-2236.

[33] Wang Z, Zhang K, Luo W, et al. HTNet for micro-expression recognition[J]. Neurocomputing, 2024, 602: 128196.

[34] Khor H Q, See J, Liong S T, et al. Dual-stream shallow networks for facial micro-expression recognition[C].2019 IEEE International Conference on Image Processing (ICIP). 2019: 36-40.

[35] Liong S T, Gan Y S, See J, et al. Shallow Triple Stream Three-dimensional CNN (STSTNet) for Micro-expression Recognition[C].2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019). 2019: 1-5.

[36] Li J, Wang T, Wang S J. Facial Micro-Expression Recognition Based on Deep Local-Holistic Network[J]. Applied Sciences, 2022, 12(9): 4643.

[37] Zhu G, Liu L, Hu Y, et al. Three-Stream Temporal-Shift Attention Network Based on Self-Knowledge Distillation for Micro-Expression Recognition[A]. arXiv, 2024.

[38] Zhou L, Mao Q, Xue L. Dual-inception network for cross-database micro-expression recognition[C]. 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019). 2019: 1-5.

[39] Hao H, Wang S, Ben H, et al. Hierarchical Space-Time Attention for Micro-Expression Recognition[A]. arXiv, 2024.

[40] Kline D M, Berardi V L. Revisiting squared-error and cross-entropy functions for training neural network classifiers[J]. Neural Computing & Applications, 2005, 14(4): 310-318.

[41] Wen Y, Zhang K, Li Z, et al. A Discriminative Feature Learning Approach for Deep Face Recognition[C].Leibe B, Matas J, Sebe N, et al. Computer Vision – ECCV 2016. Cham: Springer International Publishing, 2016: 499-515.