Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.084782

ARTICLE

DCHF: Dual-Stream Cooperative Perception with Hierarchical Fusion Network for Micro-Expression Recognition

Zishi Li and Xiaodong Huang*

Information Engineering College, Capital Normal University, Beijing, China

*Corresponding Author: Xiaodong Huang. Email: hxd@cnu.edu.cn

Received: 29 April 2026; Accepted: 04 August 2026; Published: 15 September 2026

ABSTRACT: Micro-expression recognition (MER) is a challenging task because micro-expressions are extremely short in duration, weak in intensity, and often distributed over subtle local facial regions. Existing methods either rely on handcrafted descriptors with limited representation capacity or focus on single-stream deep models that do not fully exploit structural facial dependency and cross-modal complementarity. As a result, they often fail to effectively capture subtle local muscle activations, model structured facial interactions, and robustly integrate complementary motion and appearance cues, especially under the weak-motion conditions characteristic of micro-expressions. To address these limitations, this paper proposes a Dual-Stream Cooperative Perception with Hierarchical Fusion Network (DCHF) for MER. The proposed framework jointly models temporal optical flow and spatial apex frames through a dual-stream architecture. The framework contains three key components. Adaptive Centroid-Aware Feature Extraction enhances key muscle activation regions through adaptive centroid-aware local attention while preserving global semantic context. Vertical Ipsilateral Dependency explicitly models same-side upper–lower facial coupling. The Hierarchical Dual-Stream Fusion Module progressively fuses motion and appearance information from local to global semantic levels. Extensive experiments on the Spontaneous Micro-expression Corpus (SMIC), Chinese Academy of Sciences Micro-Expression II (CASME II), Spontaneous Actions and Micro-Movements (SAMM), and their composite setting demonstrate that DCHF achieves competitive performance. In particular, on the composite benchmark, DCHF obtains 0.9121 Unweighted F1-score (UF1) and 0.9182 Unweighted Average Recall (UAR), improving the strongest compared UF1 and UAR results by 0.70 and 2.49 percentage points, respectively. These improvements suggest that adaptive centroid-aware local modeling, structure-aware ipsilateral dependency learning, and hierarchical dual-stream fusion are complementary and effective for robust MER under cross-dataset variation.

KEYWORDS: Micro-expression recognition; dual-stream learning; optical flow; apex frame; hierarchical fusion; facial dependency modeling

1 Introduction

1.1 Background and Motivation

Facial expressions are an important medium for human emotional communication. A micro-expression (ME) is a spontaneous facial movement that is extremely brief in duration, usually shorter than 500 milliseconds, weak in intensity, and not subject to conscious control [1,2]. It can reveal a person's genuine emotional state when the person attempts to suppress it, and therefore has important application value in lie detection, psychological assessment, affective computing, and security screening [3]. However, because micro-expressions are short-lived, subtle in motion, and often confined to local facial regions, micro-expression recognition (MER) remains a highly challenging task.

1.2 Key Challenges and Research Gap

Existing MER methods have attempted to address these difficulties from different perspectives. Early handcrafted approaches, such as LBP-TOP and optical-flow-based descriptors, divide the face into fixed regions and extract spatiotemporal features from each region [4–6]. Although these methods are interpretable, their representation capacity is limited and they are sensitive to alignment errors, noise, and inter-subject facial variation. Recent studies have explored graph-based relation modeling, Transformer-based local-global interaction, symmetry-aware attention, AU-guided coordination, and dual-stream learning to obtain more discriminative representations [7–11].

Nevertheless, three challenges remain central to robust MER. First, many local-region-based methods rely on fixed facial partitions, predefined landmark offsets, or heuristic ROI selection strategies. Such designs may not adapt sufficiently to inter-subject facial geometry and local muscle activation differences. Second, existing relation-modeling approaches mainly focus on generic inter-region interaction, local-global aggregation, bilateral symmetry, or AU-guided coordination. Neurophysiological and behavioral studies suggest that upper and lower facial actions may be governed by partially distinct mechanisms and can interact in concealed or mixed facial responses [12,13]. This motivates us to examine the vertical ipsilateral dependency between the brow region and the corresponding mouth corner, which has received limited explicit attention in existing region-based MER methods. Third, optical-flow-based representations are effective for subtle motion modeling but may become unstable under weak-motion or noisy conditions, whereas complementary appearance and structural cues in apex frames are often fused through shallow strategies such as direct concatenation or weighted summation.

1.3 Problem Statement and Research Objectives

Therefore, this study addresses how adaptive local representation learning, explicit ipsilateral facial dependency modeling, and hierarchical motion–appearance fusion can be jointly developed for robust MER on small-scale and class-imbalanced datasets.

To address this problem, we aim to develop a representation-learning framework that can adapt local facial modeling to subject-specific geometry, capture same-side upper–lower facial dependency without requiring complete AU annotations, and progressively integrate optical-flow motion with apex-frame appearance information. The study further evaluates whether these design goals improve MER performance under leave-one-subject-out (LOSO) evaluation using UF1 and UAR.

1.4 Proposed Framework, Research Questions, and Contributions

To address these gaps, this paper proposes a Dual-Stream Cooperative Perception with Hierarchical Fusion Network (DCHF) for micro-expression recognition. DCHF consists of three key components: Adaptive Centroid-Aware Feature Extraction (ACAFE), Vertical Ipsilateral Dependency modeling (VID), and a Hierarchical Dual-Stream Fusion Module (HDFM). ACAFE derives the centers of four salient brow and mouth-corner regions from subject-specific facial landmarks and uses these centroids to guide local attention while preserving holistic facial context. This design reduces the dependence on predefined regional centers and fixed landmark offsets, which is consistent with the localized and spatially variable nature of micro-expression cues. VID targets a different limitation of existing relation-modeling approaches. Whereas prior methods mainly emphasize generic graph relations, bilateral symmetry, or AU-level coordination, VID explicitly models same-side upper–lower dependency between brow regions and their corresponding mouth corners. This dependency is learned through directional feature interaction and structured supervision

without requiring complete AU annotations. HDFM is designed to preserve level-specific complementarity between motion and appearance representations. Rather than applying flat or single-stage fusion, it progressively integrates the two modalities at local, global, and bridge levels.

Accordingly, this study investigates three questions. First, does ACAFE improve MER performance measured by UF1 and UAR by strengthening adaptive local-global representation learning? Second, does VID provide complementary structural information, and how do its interaction and supervision components contribute to recognition? Third, does HDFM outperform flat and single-level motion–appearance fusion strategies? These questions are examined through the module-wise, VID-specific, and fusion-strategy ablation studies, respectively.

The main research contributions of this paper are summarized as follows:

- We develop ACAFE, an adaptive local-global representation mechanism that uses landmark-defined regional centroids to guide local attention while retaining holistic facial context. This reduces the dependence of local representation learning on predefined regional offsets.
- We introduce VID, which provides an explicit computational formulation of vertical ipsilateral dependency between brow regions and corresponding mouth corners. VID combines directional same-side interaction with alignment, discrimination, and cross-stream consistency supervision.
- We design HDFM to progressively integrate optical-flow motion and apex-frame appearance information at local, global, and bridge levels. The fusion ablation shows that this hierarchical design outperforms weighted summation, concatenation with an MLP, and single-level fusion variants.
- Comprehensive experiments on SMIC, CASME II, SAMM, and their composite setting validate the proposed framework. On the composite benchmark, DCHF achieves 0.9121 UF1 and 0.9182 UAR, improving the strongest compared UF1 and UAR results under this protocol by 0.70 and 2.49 percentage points, respectively. Module-wise, VID-specific, and fusion-strategy ablations further establish the contributions of the proposed components.

2 Related Work

2.1 Handcrafted Spatiotemporal Descriptors

Early studies on micro-expression recognition (MER) mainly relied on manually designed spatiotemporal descriptors to capture subtle facial motions [1,2]. A representative line of work used texture-based features, such as LBP-TOP, to encode facial dynamics on orthogonal planes [4,5], while another line employed motion-oriented descriptors, such as HOOF, MDMO, and Bi-WOOF, to model small facial movements from optical flow [6,14–16]. These methods are simple and interpretable, and for a long period they constituted the dominant technical route for MER. However, their discriminative ability depends heavily on handcrafted operators, fixed region partition strategies, and manually selected motion statistics, which makes them relatively sensitive to alignment errors, noise, and inter-subject variation. These limitations also mean that handcrafted methods usually have difficulty explicitly modeling structured dependency among facial regions.

2.2 Deep Key-Phase and Motion Representation Learning

With the development of deep learning, MER has gradually shifted from handcrafted descriptors to learnable spatiotemporal representations. Existing studies have shown that early deep MER methods mainly used the apex frame or key frames as input [6,17], and were later extended to CNN, 3D-CNN, and RNN-based architectures for modeling the onset–apex–offset process [1,2]. On this basis, multi-stream architectures further became a mainstream design, where the appearance stream captures facial texture and the motion

stream models optical flow or other dynamic cues [11,18]. This line of research is directly related to the dual-stream motion–appearance setting adopted in this paper. Nevertheless, many existing deep methods still rely on a single dominant modality or use relatively shallow interaction between streams, which limits their robustness when motion evidence is weak or noisy [1,11,18].

2.3 Local-Region and Structural Relation Modeling

Recent MER methods have increasingly focused on local facial regions rather than full-face information [7–9]. Transformer-based models have shown that self-attention can capture dependencies across frames and regions, while explicitly focusing on core muscle groups related to action units (AUs) [8,19]. A representative example is HTNet, which divides the face into four key regions and models subtle muscular movements through local attention and hierarchical aggregation [8]. Other studies have also explored graph-based or attention-based relation modeling among facial parts [7,9,20]. These studies confirm the importance of local structure, but most existing region definitions still rely on fixed partition rules, heuristic ROI selection, or predefined offsets around landmarks [7–9]. As a result, they may be less robust under inter-subject variation. In addition, although prior works have considered generic inter-region interaction, explicit modeling of vertical ipsilateral dependency between brow and mouth-corner regions remains insufficient [8–10]. These observations indicate that existing region-based methods still have limited adaptability to subject-specific facial geometry and provide insufficient modeling of targeted same-side upper–lower relations.

2.4 AU-Guided Structural Priors and Hierarchical Fusion

Another closely related research direction introduces AU information as structural prior knowledge or semantic supervision, often together with hierarchical fusion strategies. FRL-DGT performs local, global, and full-face Transformer fusion based on AU-related regions from onset–apex inputs, demonstrating the effectiveness of multi-level feature interaction for MER [19]. FED-PsyAU further emphasizes upper–lower AU coordination and combines local-to-global graph reasoning with dual-stream integration of structural features and optical flow [10]. DIANet also adopts a dual-stream design and employs cross-attention-style fusion together with consistency regularization, although it is built on phase-aware dynamic images rather than motion–appearance cooperation [21]. More recently, MER-CLIP converts AU labels into textual descriptions and aligns visual dynamics with AU semantics through vision–language learning [22]. MiHF-Tr is also related in that it emphasizes multi-information hierarchical fusion together with local/global consistency or correlation modeling [23]. These methods show that semantic priors, AU-related regions, and hierarchical feature interaction are useful for MER. However, the joint modeling of adaptive landmark-centroid local representations, explicit ipsilateral upper–lower dependency, and hierarchical motion–appearance interaction remains insufficiently explored.

2.5 Self-Supervised, Multi-Branch, and Multi-Modal Methods

Recent studies have explored self-supervised learning, pretraining, and multi-branch attention to alleviate the limited-annotation and weak-motion challenges in MER. SelfME learns motion representations through self-supervised objectives, which helps exploit subtle facial dynamics when labeled samples are scarce [24]. Micron-BERT introduces BERT-style pretraining to enhance token-level facial representation learning [25]. Dual-ATME and dual-branch cross-attention networks further strengthen local motion perception and suppress irrelevant facial responses through attention-based branch interaction [26,27]. These methods improve representation capacity and robustness from the perspectives of pretraining, self-supervision, or attention design. Nevertheless, they mainly focus on general motion representation or

branch-level attention and do not directly formulate the vertical ipsilateral dependency between brow and mouth-corner regions or the progressive local-global-bridge fusion of motion and appearance cues.

Another emerging line directly learns facial graph structures or integrates multi-scale and multi-modal representations. Direct graph-learning methods attempt to reduce the dependence on manually defined graph topology by learning facial relations from data [28]. MMTNet combines multi-modal and multi-scale Transformer representations to exploit complementary information from different sources [29], while entire-detail dual-branch modeling and Micro_NesT emphasize fine-grained detail perception and multi-scale attention for subtle local structures [30,31]. These studies demonstrate the value of structural and complementary information for MER. However, many local regions are still defined through fixed or heuristic strategies, and the specific same-side upper-lower coupling between brow and mouth-corner regions remains insufficiently investigated.

2.6 Summary and Research Gap

In summary, prior MER studies have substantially advanced handcrafted motion descriptors, deep key-phase modeling, local-region representation learning, structural relation modeling, AU-guided reasoning, and hierarchical fusion. However, three limitations remain insufficiently resolved. First, many local-region designs still depend on fixed partitions, predefined offsets, or heuristic ROI settings, which may be less adaptive to subject-specific facial geometry. Second, existing structural modeling mainly focuses on generic graph relations, bilateral symmetry, or AU coordination, while explicit vertical ipsilateral dependency between brow and mouth-corner regions remains underexplored. Third, although multi-stream methods have demonstrated the value of complementary cues, many fusion strategies remain shallow or single-level. These gaps motivate a method that jointly considers adaptive local representation, targeted ipsilateral dependency modeling, and hierarchical motion–appearance fusion.

3 Method

3.1 Overview of the Proposed Framework

Fig. 1 shows the overall architecture of DCHF. The framework contains two symmetric input streams. The motion stream takes TV-L1 optical flow and optical strain as input, while the spatial stream takes the apex RGB frame as input. Each stream first extracts local and global representations through ACAFE. The local representations are then refined by VID to model same-side upper–lower facial dependency. Finally, HDFM progressively fuses motion and appearance information at local, global, and bridge levels before classification.

3.2 Distinction from Closely Related Methods

Among recent methods, FED-PsyAU, FRL-DGT, and DIANet are among the closest to DCHF because they combine local structural reasoning with multi-stream or hierarchical interaction [10,19,21]. SOFP is also closely related, since it models optical-flow structure over four facial regions and combines structure-guided attention with local/global interaction [9]. However, DCHF differs in its technical focus. Compared with FED-PsyAU, DCHF does not rely on psychological priors or AU graphs for upper–lower coordination; instead, VID explicitly learns ipsilateral upper–lower dependency in the embedding space [10]. Compared with FRL-DGT, DCHF does not depend on a displacement generation mechanism or AU-region Transformer fusion; instead, it combines TV-L1 motion, apex appearance, and HDFM for hierarchical motion–appearance fusion [19]. Compared with DIANet, DCHF also follows a dual-stream interaction paradigm, but DIANet focuses on phase-aware dual dynamic images with consistency regularization, whereas DCHF emphasizes motion–appearance cooperation together with an explicit structural dependency

objective [21]. MER-CLIP follows a different route based on AU-to-text and vision–language alignment rather than structural dependency learning [22]. MiHF-Tr is also relevant in that it explores multi-information hierarchical fusion [23], but the main novelty of DCHF lies in the unified combination of adaptive centroid-aware local modeling, explicit ipsilateral dependency learning, and hierarchical motion–appearance fusion.

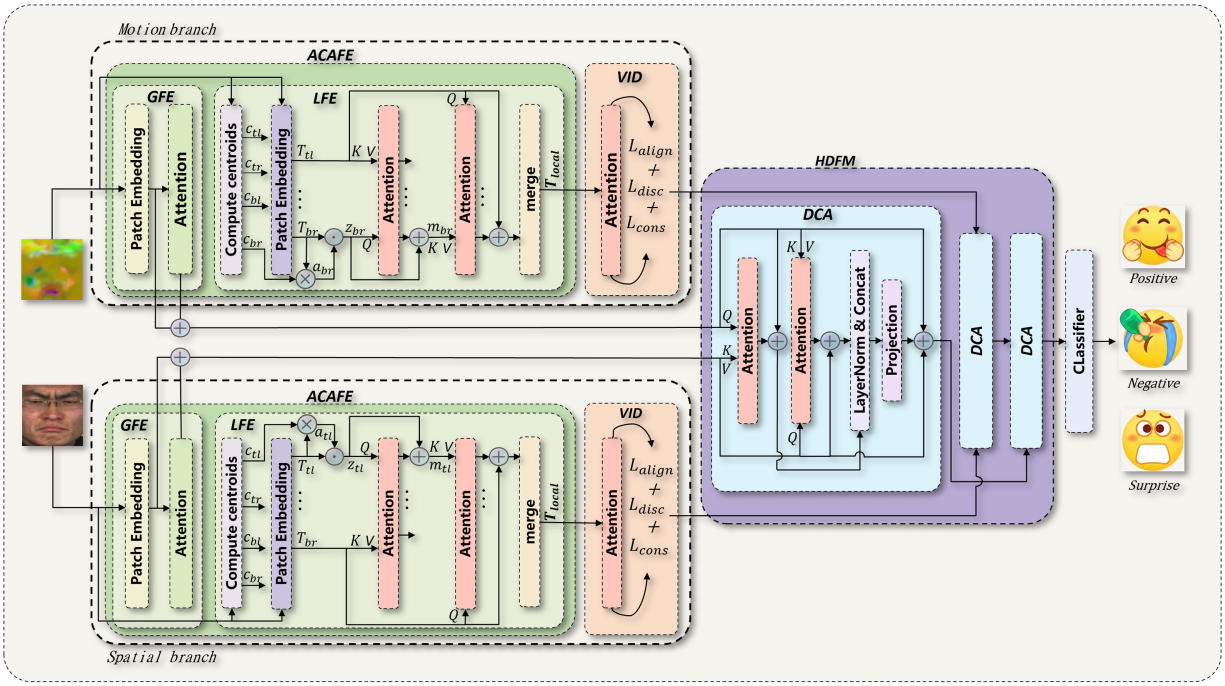

Figure 1: Overview of the proposed DCHF framework. The pipeline consists of four stages: (1) optical-flow and apex-frame inputs, (2) ACAFE-based feature extraction, (3) VID-based dependency modeling, and (4) HDFM-based fusion and classification.

3.3 Pre-Processing

A fundamental challenge in MER is that the target motion is extremely weak, whereas directly modeling full video sequences often introduces substantial temporal redundancy and irrelevant appearance variation. To obtain a compact yet discriminative motion representation, we follow the key-frame-based paradigm and construct the motion input from the onset frame and the apex frame [6,17]. Specifically, we compute the TV-L1 optical flow as [32]

$$V_m = [V_x, V_y, V_z], \tag{1}$$

where V_x and V_y denote the horizontal and vertical optical flow, respectively, and V_z denotes the optical strain [33]. Compared with raw frame differences, this representation explicitly captures both displacement and local deformation patterns, which are more suitable for subtle facial motion modeling.

For reproducibility, we describe the preprocessing pipeline used by DCHF in detail. For each sample, face detection is performed separately on the onset frame and the apex frame. The detected facial regions are cropped and resized to 224 × 224, producing normalized onset and apex faces. No additional rotation-based similarity alignment is applied; instead, the resized face crop defines a unified coordinate system for the spatial branch, landmark-based local-region construction, and cross-stream correspondence. We then

detect 68 facial landmarks on the normalized apex face and compute four regional centers corresponding to the left brow, right brow, left mouth corner, and right mouth corner.

TV-L1 optical flow is computed between the normalized onset and apex faces. Optical strain is further calculated and stacked with the horizontal and vertical flow fields to form the three-channel motion representation $V_m = [V_x, V_y, V_z]$. The optical-flow map is resized to 28×28 as the input of the motion branch. To keep the two streams spatially aligned, each regional center computed in the 224×224 apex coordinate system is mapped to the 28×28 optical-flow coordinate system by the same scale ratio. The local crop windows in the spatial and motion streams are then constructed around their corresponding regional centers. If a crop window exceeds the valid image area, the window is clipped to the image boundary so that the available facial information is preserved without introducing padded pixels.

In parallel, the normalized apex RGB frame is used as the input of the spatial branch. The rationale is that the motion branch emphasizes transient muscle dynamics, whereas the spatial branch preserves texture, structure, and appearance cues at the peak expression state. These two modalities are naturally complementary: the former captures subtle motion changes, while the latter provides stable facial structure and appearance context [11,18]. Therefore, the proposed framework adopts a dual-stream formulation to jointly model motion and appearance information before subsequent structural reasoning.

Fig. 2 illustrates the construction of the dual-stream inputs.

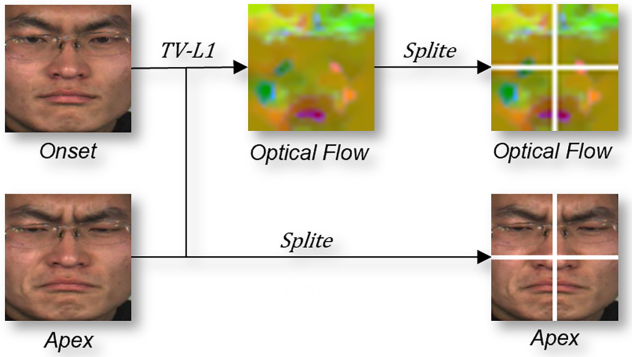

Figure 2: Construction of the dual-stream inputs.

3.4 ACAFE: Adaptive Centroid-Aware Feature Extraction

Because MER cues are subtle and spatially unstable, ACAFE jointly models local discriminative regions and global facial context [8,9].

In both the motion branch and the spatial branch, ACAFE consists of two parallel paths: a local feature extraction (LFE) path operating on four local facial regions and a global feature extraction (GFE) path operating on the original full-face image.

In the implementation, the local-region radius is set to 28 pixels in the normalized 224×224 apex coordinate system, corresponding to a 56×56 apex crop. For the motion branch, both the regional center and the local window are mapped to the 28×28 optical-flow coordinate system, resulting in a 7×7 motion crop. Each local crop is resized inside the tokenizer to 28×28 and then projected with a 7×7 patch projection into a 4×4 token grid. The token dimension is $d = 192$, all attention layers use four heads, and each head has dimension 48. The motion and spatial streams use separate tokenizers and separate ACAFE parameters.

Local feature extraction.

Given the branch input image I, we define four adaptive local regions corresponding to the left periocular region, right periocular region, left mouth corner, and right mouth corner, following the general ROI- and landmark-based local modeling paradigm in MER [7,8]. The landmark subsets are $\Omega_{tl} = \{18, 19, 37, 38\}$, $\Omega_{tr} = \{26, 27, 45, 46\}$, $\Omega_{bl} = \{49, 50, 60, 61\}$, and $\Omega_{br} = \{54, 55, 56, 65\}$. The adopted 68-point landmark indexing and the selected landmark subsets are illustrated in Fig. 3. Given the detected 68 facial landmarks $\{p_i\}$ [34], the adaptive centroid of region r is computed as

$$c_r = \frac{1}{|\Omega_r|} \sum_{i \in \Omega_r} \phi(p_i), \tag{2}$$

where $\phi(\cdot)$ maps the landmark coordinates from the image plane to the corresponding local-region coordinates. This operation yields four local crops I_{tl}, I_{tr}, I_{bl}, and I_{br}.

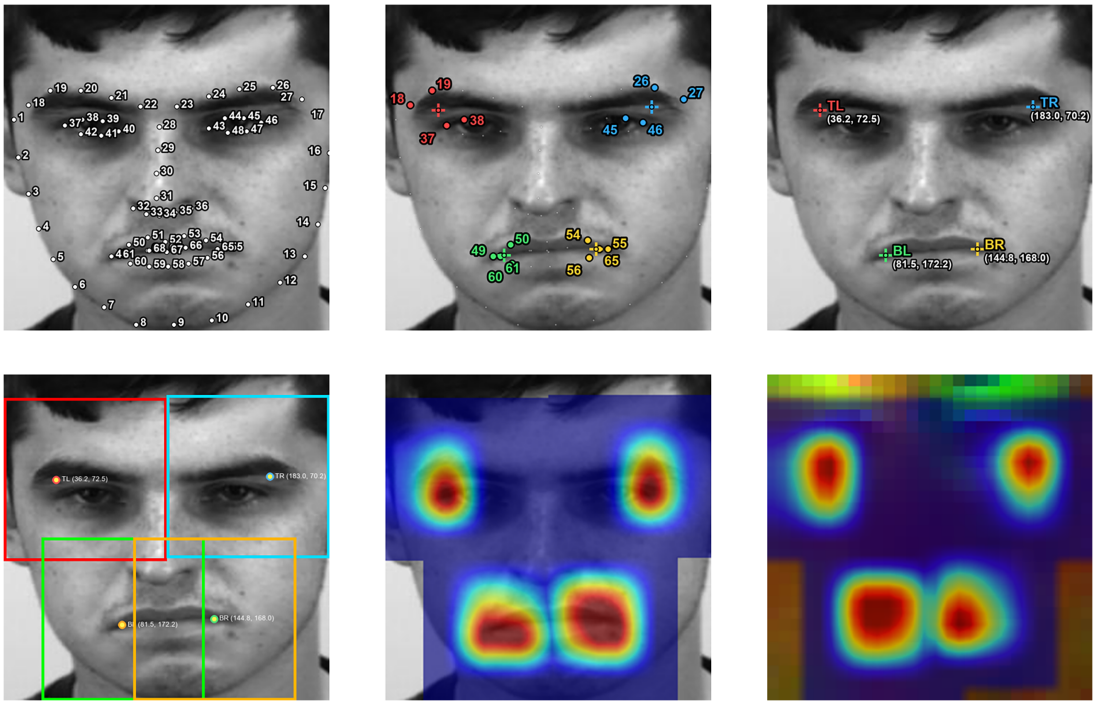

Figure 3: Region-response visualization of DCHF.

Each local crop is then converted into a token sequence [35]. Let $T_r \in R^{N_r \times d}$ denote the token sequence of region r, where N_r is the number of regional tokens. With the above setting, each facial region contains 16 tokens, i.e., $T_r \in R^{16 \times 192}$ and $T_{local} \in R^{64 \times 192}$ for each sample. The full-face path produces $T_{global} \in R^{16 \times 192}$. To emphasize the most informative region within each crop, we introduce an adaptive centroid-aware attention mechanism inspired by recent local attention and structure-aware MER designs [8,9]. Let S_r denote the set of token coordinates of region r. We assign each token location $(u, v) \in S_r$ a distance-aware weight:

$$a_r(u, v) = \frac{exp\left(-\frac{\|[u, v] - c_r\|_2^2}{\tau}\right)}{\sum_{(m,n) \in S_r} exp\left(-\frac{\|[m, n] - c_r\|_2^2}{\tau}\right)}, \tag{3}$$

where τ is a temperature parameter. The adaptive centroid token is then obtained by

$$z_r = \sum_{(u,v)\in S_r} a_r(u,v)\,T_r(u,v). \tag{4}$$

For notational simplicity, let $\mathcal{A}(\cdot)$ denote the cross-attention operator, $\mathcal{M}[\cdot]$ the region-merging operator, and $\mathcal{S}(\cdot)$ the self-attention operator.

We first use the adaptive centroid token to aggregate informative local context:

$$m_r = \mathcal{A}(Q = z_r,\ K = T_r,\ V = T_r) + z_r. \tag{5}$$

The aggregated feature is then fed back to recalibrate the entire region:

$$\hat{T}_r = T_r + \mathcal{A}(Q = T_r,\ K = m_r,\ V = m_r). \tag{6}$$

After processing all four local regions, the local representation is formed as T_{local} (Eq. (7)).

$$T_{local} = \mathcal{M}\left[\hat{T}_{tl}, \hat{T}_{tr}, \hat{T}_{bl}, \hat{T}_{br}\right]. \tag{7}$$

Global feature extraction.

The global path preserves holistic facial structure and long-range semantic context from the full-face image. Let $T_{full} \in R^{N\times d}$ denote the resulting token sequence; the global representation is

$$T_{global} = \mathcal{S}(T_{full}) + T_{full}. \tag{8}$$

This design follows the standard self-attention residual formulation used in Transformer-style visual token modeling [35,36].

ACAFE uses LayerNorm before attention and feed-forward operations. Residual connections are used after centroid aggregation, region recalibration, and global self-attention. The feed-forward layer expands the token dimension by a factor of four and then projects it back to $d = 192$. Dropout is used in the attention and feed-forward blocks to reduce overfitting.

In this way, the proposed ACAFE module jointly models subtle local motion patterns and holistic facial context through parallel LFE and GFE paths. It outputs a token-level local representation T_{local} together with a token-level global representation T_{global}. These token-level features are directly used for subsequent fusion. VID further summarizes the local tokens into region-level descriptors for explicit dependency modeling.

3.5 VID: Vertical Ipsilateral Dependency Modeling

VID is introduced to explicitly model the same-side upper–lower facial coupling that is usually missed by generic region interaction [9,10]. Fig. 4 illustrates the proposed VID module.

The module first performs same-side dependency interaction between the upper and lower facial regions on each side, and then imposes structured supervision to make the learned dependency more discriminative and cross-stream consistent.

For each branch, VID takes the local token representation $T_{local} \in R^{64\times 192}$, re-indexes it into four regional token sets T_{tl}, T_{tr}, T_{bl}, and T_{br}, where each regional token set has size 16×192, and summarizes each region as $v_r = Avg(T_r) \in R^{192}$.

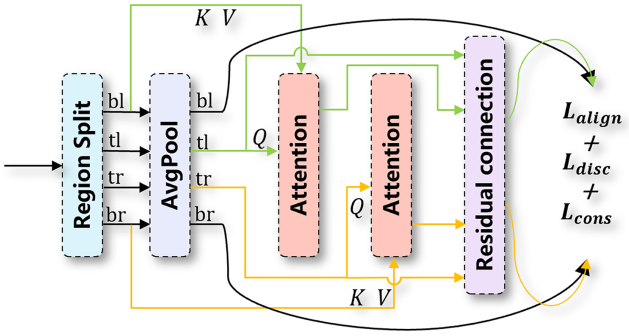

Figure 4: Illustration of the proposed VID module.

To model directional ipsilateral dependency, we construct two ipsilateral pairs (v_{tl}, T_{bl}) and (v_{tr}, T_{br}). Same-side cross-attention yields $a_L = \mathcal{A}(v_{tl}, T_{bl}, T_{bl})$ and $a_R = \mathcal{A}(v_{tr}, T_{br}, T_{br})$ [9,36], and the dependency-aware upper descriptors are

$$\tilde{v}_{tl} = v_{tl} + a_L, \qquad \tilde{v}_{tr} = v_{tr} + a_R. \tag{9}$$

This design allows the upper facial representation to selectively absorb supportive information from the corresponding lower region on the same side. As a result, the local semantics are refined through explicit same-side interaction.

(1) Alignment. The first objective enforces semantic compatibility between the upper and lower regions on the same side. With $\hat{v}_{tl} = M_L \tilde{v}_{tl}$ and $\hat{v}_{tr} = M_R \tilde{v}_{tr}$, the alignment loss is

$$L_{align} = 1 - \frac{1}{2}\left(cos(\hat{v}_{tl}, v_{bl}) + cos(\hat{v}_{tr}, v_{br})\right). \tag{10}$$

This term encourages semantically compatible same-side upper–lower representations while preserving their structural asymmetry.

(2) Discrimination. The discrimination term requires same-side similarity to be larger than cross-side similarity by a margin m:

$$L_{disc} = \frac{1}{2}\Big(max(0,\ m - cos(\hat{v}_{tl}, v_{bl}) + cos(\hat{v}_{tl}, v_{br})) + max(0,\ m - cos(\hat{v}_{tr}, v_{br}) + cos(\hat{v}_{tr}, v_{bl}))\Big). \tag{11}$$

This objective makes the learned dependency relation more selective and differentiates true ipsilateral coordination from spurious cross-side similarity.

(3) Consistency. The consistency term keeps the same structural prior compatible across the motion and spatial branches, where $s_L^{(t)}$, $s_R^{(t)}$, $s_L^{(s)}$, and $s_R^{(s)}$ denote the same-side similarities in the two modalities:

$$L_{cons} = \frac{1}{2}\left[\left(s_L^{(t)} - s_L^{(s)}\right)^2 + \left(s_R^{(t)} - s_R^{(s)}\right)^2\right]. \tag{12}$$

This term stabilizes the learned dependency structure across modalities and suppresses branch-specific structural drift, following the general idea of cross-stream consistency regularization in recent MER models [21,23].

The final VID loss is

$$L_{VID} = \alpha L_{align} + \beta L_{disc} + \gamma L_{cons}, \tag{13}$$

where α, β, and γ are trade-off coefficients. Unlike a conventional auxiliary regularization term, the proposed VID acts as a dedicated structure learning module. It jointly performs directional dependency interaction and optimization-level structural supervision. In this way, the local representations become more discriminative and structurally coherent.

VID is applied to the motion and spatial streams separately, and the left-side and right-side dependency mappings are implemented independently. The resulting VID losses are averaged. The interaction path uses residual updates, while the structured loss terms act on region-level descriptors and therefore add only limited computational overhead.

3.6 HDFM: Hierarchical Dual-Stream Fusion Module

HDFM is introduced to preserve level-specific complementarity between motion and appearance features, which may be insufficiently captured by flat or single-stage fusion strategies [8,19,23].

Let $T_{local}^{(t)}$ and $T_{local}^{(s)}$ denote the local token-level features from the motion and spatial branches, respectively. We first perform local-level fusion to align subtle motion patterns with corresponding appearance cues:

$$F_{local} = DCA\left(T_{local}^{(t)},\ T_{local}^{(s)}\right) + \frac{T_{local}^{(t)} + T_{local}^{(s)}}{2}. \tag{14}$$

This stage focuses on fine-grained complementary learning, where the motion stream highlights subtle muscular changes and the spatial stream supplements them with stable structural and texture context.

Similarly, let $T_{global}^{(t)}$ and $T_{global}^{(s)}$ denote the global token-level features from the two branches. We further perform global-level fusion as

$$F_{global} = DCA\left(T_{global}^{(t)},\ T_{global}^{(s)}\right) + \frac{T_{global}^{(t)} + T_{global}^{(s)}}{2}. \tag{15}$$

Unlike local fusion, this stage emphasizes the integration of holistic semantic information and long-range facial dependencies. It complements localized motion modeling with broader contextual understanding.

After obtaining the local and global fused representations, we further perform a bridge fusion to explicitly model local–global interaction:

$$F_{fusion} = DCA\left(F_{local},\ F_{global}\right). \tag{16}$$

In this way, HDFM follows a fine-to-coarse yet semantically disentangled fusion paradigm inspired by prior hierarchical fusion designs in MER [8,19,23]. It first preserves subtle local motion cues, then incorporates global structural context, and finally bridges the two levels into a single discriminative representation.

The core fusion operator, namely dual cross-attention (DCA), mutually updates the two inputs through bidirectional cross-attention, concatenates the residual-enhanced streams after layer normalization, and projects them back to the original embedding dimension [36,37].

In our implementation, the local inputs to HDFM have size $B \times 64 \times 192$, while the global inputs have size $B \times 16 \times 192$. The local, global, and bridge fusion stages use independent DCA blocks. Each DCA block contains two directional cross-attention operations with four heads. In the local and global DCA blocks, the residual-enhanced streams are concatenated along the channel dimension to $B \times N \times 384$, normalized, and projected back to $B \times N \times 192$, where $N = 64$ for local fusion and $N = 16$ for global fusion. The bridge stage operates on the pooled local and global descriptors with size $B \times 1 \times 192$. A feed-forward residual block is applied before classification.

3.7 Classification and Optimization

The optimization objective combines classification, cross-modal alignment, and VID supervision in a unified training loss [21,23].

$$L = L_{cls} + \lambda_{local} L_{local} + \lambda_{global} L_{global} + \lambda_{vid} L_{VID}, \tag{17}$$

where L_{cls} denotes the classification loss. L_{local} and L_{global} denote the local-level and global-level cross-modal alignment losses, respectively. L_{VID} denotes the proposed vertical ipsilateral dependency objective. In practice, L_{local} is implemented as token-wise cosine alignment between the two branches. L_{global} imposes cosine consistency on their global descriptors [21,23]. By jointly optimizing these objectives, the model is encouraged to learn discriminative, cross-modally aligned, and structurally coherent representations for MER.

4 Experiments

4.1 Datasets and Evaluation Protocol

The experimental evaluation is designed to verify whether DCHF can improve MER performance under class-imbalanced, subject-independent, small-sample, and cross-dataset conditions. An experimental setup is suitable for this purpose because MER models must be evaluated on annotated video samples with fixed class labels, controlled train-test separation, and directly comparable quantitative metrics. We therefore adopt leave-one-subject-out (LOSO) evaluation to prevent subject identity leakage and to test generalization to unseen subjects. In addition, we report both single-dataset evaluation and composite-dataset evaluation, so that the model can be assessed under dataset-specific conditions and under stronger cross-dataset variation.

We evaluate the proposed method on three widely used spontaneous micro-expression benchmarks, namely SMIC, CASME II, and SAMM, as well as their composite setting [38–40]. These datasets are suitable for evaluating MER models because they contain spontaneous micro-expression samples with annotated key frames and represent different acquisition conditions, frame rates, subject groups, and motion characteristics. Specifically, SMIC represents a relatively low-frame-rate and weak-motion setting, CASME II provides higher-frame-rate samples under relatively consistent acquisition conditions, and SAMM contains greater inter-subject and demographic diversity. Their combination therefore enables evaluation under complementary motion, subject, and acquisition conditions. Following the standard MEGC protocol, all samples are grouped into three emotion categories, i.e., negative, positive, and surprise [41]. The composite setting, denoted as Full, merges SMIC, CASME II, and SAMM to evaluate cross-dataset generalization [41,42], while the single-dataset evaluation (SDE) protocol reports results on each dataset individually.

Table 1 summarizes the basic statistics of the filtered three-class subsets used under the MEGC-style evaluation protocol.

Table 1: Basic statistics of the filtered three-class subsets used under the MEGC-style evaluation protocol.

Database	SAMM	CASME II	SMIC
Subjects	28	24	16
Samples	133	145	164
Frame rate	200	200	100
Negative	92	88	70
Positive	26	32	51
Surprise	15	25	43

To ensure fair comparison under the highly imbalanced class distribution of MER datasets, we report the Unweighted F1-score (UF1) and Unweighted Average Recall (UAR) as the main evaluation metrics [41]. Compared with overall accuracy, UF1 and UAR assign equal importance to each emotion category and are less sensitive to class-frequency bias. They are therefore more suitable for evaluating MER performance under class-imbalanced conditions.

Specifically, UAR is defined as the average recall over all C classes:

$$UAR = \frac{1}{C}\sum_{i=1}^{C}\frac{TP_i}{TP_i + FN_i}, \tag{18}$$

where TP_i and FN_i denote the true positives and false negatives of the i-th class, respectively.

For each class, the F1-score is computed as

$$F1_i = \frac{2P_iR_i}{P_i + R_i}, \tag{19}$$

where P_i and R_i denote the precision and recall of the i-th class. The UF1 is then obtained by averaging the class-wise F1-scores:

$$UF1 = \frac{1}{C}\sum_{i=1}^{C}F1_i. \tag{20}$$

Here, precision and recall are defined as

$$P_i = \frac{TP_i}{TP_i + FP_i}, \qquad R_i = \frac{TP_i}{TP_i + FN_i}, \tag{21}$$

where FP_i denotes the false positives of the i-th class.

4.2 Implementation Details

All experiments are conducted on a single NVIDIA RTX 3090 GPU under Ubuntu 24.04 with PyTorch 2.2.2 [43]. For each sample, the onset and apex frames are first processed by face detection and cropping, and the cropped face regions are resized to 224 × 224. We compute the TV-L1 optical flow between the normalized onset and apex frames as the motion input, while the normalized apex RGB frame is used as the spatial input. In addition, 68 facial landmarks are detected on the normalized apex frame to construct the adaptive regional centroids required by ACAFE. Each local crop is constructed with a radius of 28 pixels in the 224 × 224 apex coordinate system. If the crop window exceeds the image boundary, it is clipped to the valid image area. The optical-flow map is resized to 3 × 28 × 28, and the apex RGB input is resized to 3 × 224 × 224. The

regional centers obtained from the 224 × 224 apex coordinate system are mapped to the 28 × 28 optical-flow coordinate system to maintain spatial correspondence between the two streams. The patch size is 7 × 7, the token dimension is 192, and all attention blocks use four heads. The model is trained end-to-end under the LOSO protocol using the same preprocessing pipeline and training configuration across all datasets, unless otherwise specified.

During training, the motion and spatial branches are jointly optimized in a unified dual-stream framework. ACAFE first produces branch-specific local and global token representations, VID further imposes structure-aware regularization on the local tokens, and HDFM then fuses the local and global features across the two modalities. The classification objective, cross-modal alignment losses, and the proposed VID regularization are optimized simultaneously. To ensure fair comparison, all ablation settings share the same experimental environment, data split protocol, and optimization strategy. Unless otherwise stated, the reported results are obtained from the final converged model under the corresponding evaluation setting. No information from the held-out test subject is used for model selection under the LOSO protocol.

The optimizer is AdamW [44], the initial learning rate is 1×10^{-3}, the weight decay is 1×10^{-4}, the batch size is 64, and the model is trained for 150 epochs. A cosine annealing warm-restart schedule is used with a minimum learning rate of 1×10^{-6}. The dropout rate is set to 0.1. The loss weights are set to $\lambda_{local} = 0.2$, $\lambda_{global} = 0.2$, and $\lambda_{vid} = 0.3$. Fixed random seeds of 3407, 3408, and 3409 are used for reproducibility checks. To report the computational cost, DCHF contains 6.615M parameters, requires 1.772 GFLOPs for a single forward pass, and achieves an average inference time of 16.189 ms per sample on the RTX 3090 GPU under the default input sizes.

Hyperparameter selection.

The key hyperparameters are selected according to the input scale, token resolution, and the relative roles of different loss terms, and are kept fixed across all datasets and ablation settings. The local-region radius is set to 28 pixels in the normalized 224 × 224 apex coordinate system, producing a 56 × 56 crop. This setting covers the landmark-defined brow and mouth-corner neighborhoods while avoiding excessive background or unrelated facial regions. After mapping to the 28 × 28 optical-flow map, the corresponding motion crop becomes 7 × 7, which is consistent with the patch size used by the tokenizer.

The ACAFE temperature parameter is set to $\tau = 1.65$. This parameter controls the sharpness of the centroid-aware spatial weighting. A moderate value is used to emphasize tokens close to the adaptive centroid while preventing the attention from collapsing to a single token on the 4 × 4 regional token grid. For the loss weights, $\lambda_{local} = 0.2$, $\lambda_{global} = 0.2$, and $\lambda_{vid} = 0.3$ are used. The local and global alignment losses are assigned equal weights because they regularize complementary cross-modal consistency at different semantic levels. The VID loss is assigned a slightly larger but still auxiliary weight to strengthen structural dependency learning without overwhelming the primary classification objective.

4.3 Comparison with State-of-the-Art

Table 2 reports the comparison with representative handcrafted and deep learning methods under both the composite and single-dataset protocols. The baseline results are directly cited from the corresponding published papers rather than reproduced in our codebase. To ensure a meaningful comparison, we include methods that report results under the same MEGC-style three-class label mapping, LOSO evaluation protocol, and UF1/UAR metrics whenever available. Since implementation details such as face preprocessing, augmentation, and backbone configuration may still differ across published methods, the SOTA comparison

should be interpreted as a protocol-level comparison, while the ablation studies in Section 4.4 provide controlled evidence under the same implementation setting.

Table 2: Comparison with representative handcrafted and deep learning methods under the composite and single-dataset evaluation protocols. The best result in each column is shown in bold, and the second-best result is underlined.

Methods	Full		SMIC		CASME II		SAMM	
	UF1	UAR	UF1	UAR	UF1	UAR	UF1	UAR
AlexNet (2012) [45]	0.6933	0.7154	0.6201	0.6373	0.7994	0.8312	0.6104	0.6642
LBP-TOP [4,5]	0.5882	0.5785	0.2000	0.5280	0.7026	0.7429	0.3954	0.4102
Bi-WOOF (2018) [6]	0.6296	0.6227	0.5727	0.5829	0.7805	0.8026	0.5211	0.5139
OFF-ApexNet (2019) [17]	0.7196	0.7096	0.6817	0.6695	0.8764	0.8681	0.5409	0.5392
JGULF (2024) [46]	0.8482	0.8460	0.8120	0.8143	0.8841	0.8765	0.8337	0.8192
HTNet (2024) [8]	0.8603	0.8475	0.8049	0.7905	0.9532	0.9516	0.8131	0.8124
HFA-Net (2025) [47]	0.8800	0.8660	0.7786	0.7786	0.9738	0.9754	**0.9002**	**0.8938**
EDMDBN (2025) [30]	0.8821	0.8933	0.7948	0.8085	0.9484	0.9619	0.8336	0.8661
MERba (2025) [48]	0.8840	0.8777	**0.9498**	**0.9470**	0.8701	0.8423	0.8312	0.7639
LVTEA (2025) [49]	0.8839	0.8700	0.8203	0.8137	0.9676	0.9613	0.8392	0.8306
SODA4MER (2025) [50]	–	–	0.8881	0.8863	0.8870	0.8809	–	–
CAUT (2025) [51]	0.8800	0.8833	0.8340	0.8351	0.9689	0.9688	0.8306	0.8402
ME-NAS (2026) [52]	0.8263	0.8472	0.6717	0.7123	0.8913	0.8942	0.8635	0.8873
FPS (2026) [53]	–	–	0.8535	–	0.8462	–	0.8473	–
DSBICNet (2026) [54]	0.8812	0.8900	0.8249	0.8297	**0.9928**	0.9896	0.8670	0.8641
LKDTNet (2026) [55]	0.9051	0.8914	0.8315	0.8384	0.9601	0.9583	0.8613	0.8177
DCHF (Ours)	**0.9121**	**0.9182**	0.9063	0.9090	0.9845	**0.9924**	0.8120	0.8216

On the composite setting, the proposed DCHF achieves the best overall performance, yielding 0.9121 UF1 and 0.9182 UAR. Compared with the strongest compared results listed in Table 2 under this protocol, DCHF improves UF1 by 0.70 percentage points over LKDTNet and UAR by 2.49 percentage points over EDMDBN. This improvement is notable because the composite protocol is substantially more challenging than single-dataset evaluation: it mixes SMIC, CASME II, and SAMM, which differ in frame rate, subject diversity, and data distribution, and therefore places greater demands on feature robustness and cross-dataset generalization.

On the individual datasets, DCHF exhibits different behaviors. On SMIC, DCHF achieves 0.9063 UF1 and 0.9090 UAR, which is competitive but not the best result in Table 2. This may be mainly attributed to the fact that SMIC is relatively low-frame-rate and contains weaker motion patterns, making subtle discriminative cues harder to preserve. Under this setting, the proposed ACAFE and hierarchical dual-stream fusion remain beneficial, as they explicitly emphasize subtle regional motion while preserving complementary appearance context, which helps maintain robust performance when motion evidence is weak.

On CASME II, DCHF achieves 0.9845 UF1 and 0.9924 UAR. Although the UF1 is slightly lower than the best competing result, DCHF obtains the best UAR, indicating stronger recall balance across categories. This suggests that our model is effective at maintaining balanced recognition across categories on a relatively clean and high-consistency dataset. At the same time, the very strong UF1 of competing methods such as DSBICNet is also understandable, since their dynamic-static bidirectional interaction and saliency-guided

extraction are highly effective when annotations are consistent and the signal-to-noise ratio is favorable. Therefore, on CASME II, DCHF should be viewed as highly competitive rather than uniformly dominant. Its main advantage appears to lie in more balanced recall.

On SAMM, DCHF achieves 0.8120 UF1 and 0.8216 UAR, which is lower than the best-performing methods. This may be mainly because SAMM contains stronger inter-subject and cross-ethnicity variation, making stable region alignment and subtle cue extraction more difficult. Existing works such as MERba [48] explicitly note that stronger performance on SAMM is closely related to better generalization under diverse ethnic groups, while DSBICNet [54] also attributes its relative difficulty on SAMM to the higher diversity of the dataset. In our case, although ACAFE is effective in general, the same structural prior may become less stable under stronger appearance and morphology variation, which partially limits performance on SAMM. This also indicates that SAMM remains the most challenging benchmark for evaluating the robustness of MER models.

The advantage of DCHF appears most clearly on the composite dataset. In contrast to single-dataset evaluation, the composite protocol amplifies the domain gap between datasets and therefore penalizes methods that rely too heavily on a single representation or a flat fusion strategy. Prior works have shown that simple local/global concatenation or one-stage fusion often fails to fully exploit complementarity between feature streams, while methods designed primarily for one dataset may not transfer well under cross-dataset variation. The performance of DCHF on this setting is consistent with the hypothesis that it benefits from three aspects. ACAFE enhances subtle facial responses around adaptive landmark-defined regional centroids. Vertical ipsilateral dependency modeling introduces an additional structural prior beyond independent regional encoding. The hierarchical dual-stream fusion module performs progressive interaction between motion and appearance features at local and global levels. Together, these components may improve robustness to dataset shift.

Fig. 3 provides a qualitative region-response visualization of DCHF. The highlighted responses are mainly located around the brow and mouth-corner neighborhoods, which is consistent with the proposed adaptive local-region construction and same-side upper–lower dependency modeling. This visualization provides complementary qualitative evidence for the region-focused behavior of DCHF, while the controlled ablation studies in Section 4.4 provide quantitative evidence for the contributions of ACAFE, VID, and HDFM.

Fig. 5 further illustrates the confusion matrices across the composite setting and the three individual datasets.

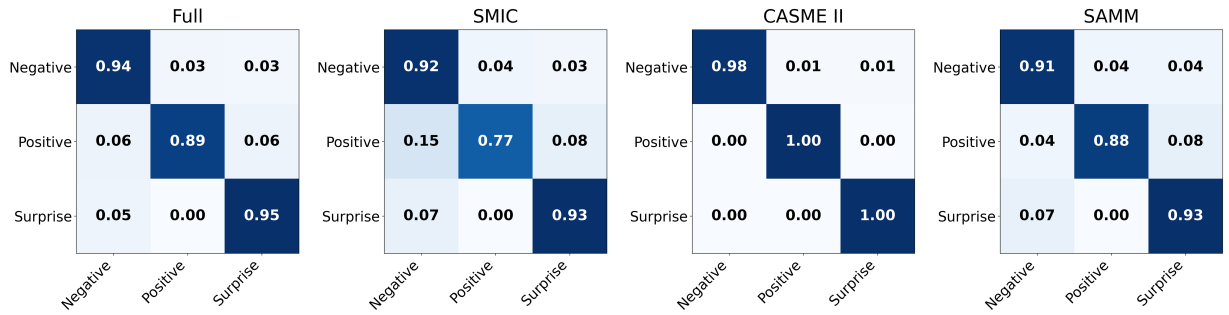

Figure 5: Confusion matrices of DCHF on the composite setting and the three individual datasets.

The confusion matrices provide a class-wise view of the recognition behavior and complement the quantitative comparison reported in Table 2. In the composite setting, the model maintains robust diagonal

responses for all categories, achieving recognition rates of 0.94, 0.89, and 0.95 for negative, positive, and surprise, respectively, further supporting its generalization ability under cross-dataset variability. On the high-consistency CASME II dataset, DCHF achieves very high class-wise recognition rates, reaching 0.98 for the negative class and 1.00 for both positive and surprise classes. Although the diverse SAMM benchmark and the low-frame-rate SMIC present greater challenges, particularly in separating the positive class, the model still maintains stable and competitive accuracy for negative and surprise expressions. Overall, these results suggest that DCHF provides robust and structurally coherent representations that align well with the intrinsic characteristics of micro-expressions.

4.4 Ablation Studies

To further validate the effectiveness of the proposed design, we conduct ablation studies from three perspectives: (1) module-wise ablation to quantify the independent contribution of each major component, (2) VID-specific ablation to examine the effect of its same-side interaction path and structured regularization terms, and (3) fusion-strategy ablation to verify whether the proposed hierarchical dual-stream fusion is more effective than simpler alternatives. Unless otherwise stated, all ablation experiments are conducted under the same LOSO protocol and training configuration as the full model, and the results are reported on the composite setting in terms of UF1 and UAR.

4.4.1 Module-Wise Ablation

We first perform a leave-one-out module ablation on the composite setting to quantify the contribution of each major component in DCHF. Starting from the full model, we remove one core module at a time while keeping the remaining architecture and training protocol unchanged. Specifically, we compare the following settings: (1) DCHF-Base, a lightweight dual-stream baseline without ACAFE, VID, or HDFM; (2) DCHF w/o ACAFE, where adaptive centroid-aware local–global representation learning is removed; (3) DCHF w/o VID, where the ipsilateral dependency modeling module is removed; (4) DCHF w/o HDFM, where hierarchical dual-stream fusion is replaced by simple concatenation followed by an MLP; and (5) the complete DCHF.

Table 3 summarizes the module-wise ablation results.

Table 3: Module-wise ablation study on the composite setting. ✓ and ✗ denote whether the corresponding component is enabled. UF1 denotes Unweighted F1-score, and UAR denotes Unweighted Average Recall.

Setting	ACAFE	VID	HDFM	UF1	UAR
DCHF-Base	✗	✗	✗	0.8473	0.8391
w/o ACAFE	✗	✓	✓	0.8842	0.8815
w/o VID	✓	✗	✓	0.9042	0.9081
w/o HDFM	✓	✓	✗	0.8972	0.9015
Full DCHF	✓	✓	✓	0.9121	0.9182

Removing any major component leads to a clear performance drop, confirming that ACAFE, VID, and HDFM are all beneficial to the final recognition performance. Among them, removing ACAFE causes the largest degradation, reducing UF1 from 0.9121 to 0.8842 and UAR from 0.9182 to 0.8815. This result suggests that adaptive centroid-aware local modeling together with global structural context learning makes the largest contribution among the evaluated components, since the quality of downstream dependency modeling and cross-modal fusion strongly depends on robust feature extraction.

Removing HDFM also produces a notable decline, with UF1 and UAR dropping to 0.8972 and 0.9015, respectively. This suggests that the advantage of DCHF does not come merely from using two modalities, but from hierarchically integrating motion and appearance information across semantic levels. By contrast, removing VID results in a smaller but still consistent decrease to 0.9042 UF1 and 0.9081 UAR, showing that explicit same-side upper–lower dependency modeling provides complementary structural supervision beyond representation learning and fusion. Finally, the large gap between DCHF-Base and the full DCHF further indicates that the gains are not produced by any single isolated technique, but by the coordinated effect of adaptive feature extraction, structure-aware dependency modeling, and hierarchical dual-stream fusion.

4.4.2 VID-Specific Ablation

Since VID is one of the key contributions of this work, we further analyze how its individual components contribute to the final performance. According to the current formulation in Section 3, VID consists of a same-side forward interaction path together with three regularization terms, namely alignment, discrimination, and consistency. We therefore compare six variants: (1) w/o VID; (2) w/o the interaction path; (3) w/o L_{align}; (4) w/o L_{disc}; (5) w/o L_{cons}; and (6) the complete VID.

Table 4 reports the VID-specific ablation results.

Table 4: VID-specific ablation study on the composite setting. ✓ and ✗ indicate whether the corresponding VID component is enabled.

Setting	Interaction	L_{align}	L_{disc}	L_{cons}	UF1	UAR
w/o VID	✗	✗	✗	✗	0.9042	0.9081
w/o Interaction	✗	✓	✓	✓	0.9055	0.9094
w/o L_{align}	✓	✗	✓	✓	0.9078	0.9128
w/o L_{disc}	✓	✓	✗	✓	0.9071	0.9120
w/o L_{cons}	✓	✓	✓	✗	0.9094	0.9147
Full VID	✓	✓	✓	✓	0.9121	0.9182

As shown in Table 4, removing any part of VID leads to lower performance than the full setting, confirming that same-side interaction and structured supervision are both beneficial. Removing the interaction path causes the largest drop among the partial variants, reducing UF1/UAR from 0.9121/0.9182 to 0.9055/0.9094. This indicates that the optimization terms alone cannot fully replace explicit same-side feature exchange during forward propagation.

Among the three regularization terms, removing L_{cons} is the least harmful, while removing L_{align} or L_{disc} causes a more obvious decline. In particular, the drop is slightly larger when L_{disc} is removed than when L_{align} is removed, suggesting that semantic alignment and cross-side discrimination are the main drivers of dependency-aware local modeling, whereas consistency mainly stabilizes the learned structure across the two modalities. The full VID yields the best UF1 and UAR, which confirms that forward interaction, alignment, discrimination, and consistency are complementary within the proposed design.

4.4.3 Fusion-Strategy Ablation

We further compare the proposed HDFM against several commonly used fusion strategies to verify whether hierarchical dual-stream fusion is indeed necessary for MER. Specifically, we consider: (1) direct

concatenation followed by an MLP; (2) weighted summation; (3) local-only fusion without the global stage; (4) global-only fusion without the local stage; and (5) the complete hierarchical HDFM with local fusion, global fusion, and bridge fusion.

Table 5 reports the fusion-strategy ablation results.

Table 5: Fusion-strategy ablation study on the composite setting. ✓ and ✗ denote whether the corresponding fusion stage is enabled.

Setting	Local Fusion	Global Fusion	Bridge Fusion	Operator	UF1	UAR
Weighted Sum	✗	✗	✗	Sum	0.8926	0.8961
Concat + MLP	✗	✗	✗	Concat/MLP	0.8972	0.9015
Global-only Fusion	✗	✓	✗	DCA	0.8994	0.9038
Local-only Fusion	✓	✗	✗	DCA	0.9048	0.9089
Full HDFM	✓	✓	✓	DCA	0.9121	0.9182

Direct weighted summation performs the worst, while concatenation followed by an MLP yields a moderate improvement. This indicates that learnable fusion is preferable to naive feature aggregation, but flat fusion strategies are still insufficient to fully exploit motion–appearance complementarity for MER. Among the hierarchical variants, global-only fusion improves over the flat baselines, suggesting that holistic semantic context is useful for cross-modal interaction. However, local-only fusion further outperforms global-only fusion, which is consistent with the fact that the most discriminative cues in MER are often concentrated in subtle local muscle activations.

The complete DCHF achieves the best performance, reaching 0.9121 UF1 and 0.9182 UAR. It surpasses local-only fusion by a clear margin, indicating that the final gain does not come from adding a single fusion stage alone. Instead, it likely comes from the coordinated effect of local fusion, global fusion, and bridge-level interaction. This observation is consistent with the nature of MER, where micro-expressions are simultaneously local and structured: discriminative cues are often localized in small facial regions, yet their interpretation also depends on broader facial context.

4.4.4 Summary of Ablation Findings

Overall, the ablation study supports three main conclusions. First, ACAFE, VID, and HDFM each contribute positively to the final performance, with ACAFE providing the largest gain, HDFM the second largest gain, and VID further offering stable complementary improvement. Second, the current VID formulation is most effective when the same-side interaction path and the three regularization terms are jointly enabled rather than partially removed. Third, the advantage of DCHF comes not merely from using two modalities, but from using a hierarchy-aware fusion strategy that explicitly separates local interaction, global interaction, and bridge-level integration. Taken together, these observations suggest that the proposed method is not a loose combination of several techniques, but a coherent framework in which ACAFE, structured dependency regularization, and hierarchical motion–appearance fusion work together to improve MER.

5 Discussion

The above results indicate that the main strength of DCHF lies in its robustness under heterogeneous evaluation settings, especially on the composite benchmark. This suggests that the proposed framework does not merely fit a specific dataset bias, but learns a more transferable representation for MER. From a modeling perspective, this advantage arises from the cooperative effect of the three main components: ACAFE

emphasizes subtle facial responses around adaptive landmark-defined centroids, VID introduces additional ipsilateral structural supervision, and HDFM progressively aligns motion and appearance information from local to global levels. Their combination enables the model to preserve both fine-grained motion sensitivity and broader semantic stability, which is particularly beneficial in cross-dataset scenarios. The ablation results further support that the effectiveness of DCHF comes from the coordinated interaction of these modules rather than from a single isolated component.

The experimental findings also directly address the guiding questions posed in the Introduction. The first question concerns the contribution of adaptive centroid-aware local-global modeling. As shown in Table 3, removing ACAFE causes the largest performance reduction among the three major modules, indicating that adaptive regional representation is important for MER performance measured by UF1 and UAR. The second question concerns vertical ipsilateral dependency. Tables 3 and 4 show that VID provides a consistent complementary gain, and that both the same-side interaction path and the structured supervision terms contribute to the final performance. The third question concerns hierarchical fusion. Table 5 demonstrates that the complete HDFM outperforms weighted summation, concatenation with an MLP, and single-level fusion variants, confirming the benefit of progressive local, global, and bridge-level interaction.

At the same time, the improvement is not uniform across all datasets. In particular, SAMM remains more challenging, indicating that stronger subject diversity and appearance variation can still weaken the stability of the learned structural prior. This suggests that, although DCHF already captures meaningful local and cross-modal structure, its robustness to cross-subject variation can be further improved. Recent MER research is increasingly moving toward multi-modal fusion, multi-scale attention, direct graph learning, and robustness under less controlled or more realistic visual conditions [28,29,31,56,57]. Future work may therefore consider richer temporal modeling, more adaptive structural priors, and more robust fusion mechanisms for ambiguous or low-resource MER settings.

6 Threats to Validity

Several threats to validity should be considered.

First, the external validity of the results is limited by the scale and diversity of existing MER datasets. Although SMIC, CASME II, SAMM, and their composite setting cover different frame rates, acquisition conditions, and subject groups, they are still relatively small-scale laboratory datasets. Therefore, the generalization ability of DCHF to in-the-wild scenarios, long unconstrained videos, or alternative emotion taxonomies requires further validation.

Second, the internal validity of the proposed framework depends on the reliability of the preprocessing pipeline. DCHF relies on facial landmarks, onset–apex optical flow, and apex RGB frames as inputs. Errors in landmark localization, face alignment, apex-frame selection, or optical-flow estimation may affect the quality of local region construction and the stability of adaptive centroid-aware modeling. This limitation becomes more significant under occlusion, low image quality, or extremely weak motion conditions.

Third, the comparison validity may be influenced by differences in implementation details across published methods. Although we follow the standard LOSO protocol and MEGC-style three-class setting, some baseline results are directly cited from prior work rather than reproduced under a unified codebase. Variations in preprocessing, data augmentation, backbone architectures, and hyperparameter tuning may therefore affect absolute performance comparisons. For this reason, the ablation studies under the same implementation setting are used as the main evidence to verify the contributions of ACAFE, VID, and HDFM.

Fourth, the construct validity is influenced by the choice of evaluation metrics and label taxonomy. UF1 and UAR are adopted to reduce the effect of class imbalance, but they may not fully capture temporal dynamics, emotion intensity, or fine-grained differences among visually similar micro-expression categories.

7 Conclusion

This paper presented DCHF, a dual-stream cooperative perception framework for micro-expression recognition that jointly models temporal optical flow and spatial apex frames in a structured and hierarchical manner. The proposed method integrates ACAFE for adaptive centroid-aware local and global representation learning, VID for explicit ipsilateral dependency modeling, and HDFM for progressive motion–appearance fusion across semantic levels. In this way, DCHF addresses weak local motion cues and the difficulty of effectively integrating complementary modalities under small-scale and imbalanced data conditions.

The quantitative results highlight the practical significance of the proposed design. On the composite benchmark, DCHF achieves 0.9121 UF1 and 0.9182 UAR, improving the strongest compared UF1 and UAR results listed in Table 2 by 0.70 and 2.49 percentage points, respectively. The confusion-matrix analysis and ablation studies further indicate that this gain is not caused by a single isolated component, but by the combined effect of adaptive centroid-aware local modeling, structure-aware ipsilateral dependency learning, and hierarchical dual-stream fusion. These findings suggest that explicitly coupling local facial structure with progressive motion–appearance interaction is an effective direction for robust MER under cross-dataset variation.

Several limitations remain. DCHF shows its clearest advantage on the composite benchmark, but it is not uniformly dominant on every single dataset, especially on SAMM, where larger inter-subject and demographic variation may reduce the stability of landmark-centered local modeling. In addition, the framework still depends on reliable landmark localization and optical-flow estimation, which may be affected by low image quality, occlusion, or extremely weak motion. Future work will therefore focus on more robust region localization, domain-adaptive fusion, and validation under less controlled or more realistic MER scenarios. This direction is also consistent with recent reviews that highlight a broader shift from handcrafted descriptors toward pretraining, graph learning, and multi-branch fusion in MER [58,59].

Acknowledgement: Not applicable.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: Study conception: Zishi Li; Draft preparation: Zishi Li; Writing, review, and editing: Xiaodong Huang. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The datasets used in this study are publicly available from their original providers subject to their access policies. The processed experimental settings and implementation details are available from the corresponding author upon reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Ben X, Ren Y, Zhang J, Wang SJ, Kpalma K, Meng W, et al. Video-based facial micro-expression analysis: a survey of datasets, features and algorithms. IEEE Trans Pattern Anal Mach Intell. 2022;44(9):5826–46.

2. Pan H, Xie L, Wang Z, Liu B, Yang M, Tao J. Review of micro-expression spotting and recognition in video sequences. Virtual Real Intell Hardw. 2021;3(1):1–17. doi:10.1016/j.vrih.2020.10.003.
3. Yildirim S, Chimeumanu MS, Rana ZA. The influence of micro-expressions on deception detection. Multimed Tools Appl. 2023;82(19):29115–33. doi:10.1007/s11042-023-14551-6.
4. Zhao G, Pietikäinen M. Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Trans Pattern Anal Mach Intell. 2007;29(6):915–28. doi:10.1109/tpami.2007.1110.
5. Wang Y, See J, Phan RCW, Oh YH. Efficient spatio-temporal local binary patterns for spontaneous facial micro-expression recognition. PLoS One. 2015;10(5):e0124674. doi:10.1371/journal.pone.0124674.
6. Liong ST, See J, Wong KS, Phan RCW. Less is more: micro-expression recognition from video using apex frame. Signal Process Image Commun. 2018;62:82–92.
7. Wei J, Peng W, Lu G, Li Y, Yan J, Zhao G. Geometric graph representation with learnable graph structure and adaptive AU constraint for micro-expression recognition. IEEE Trans Affect Comput. 2024;15(3):1343–57. doi:10.1109/taffc.2023.3340016.
8. Wang Z, Zhang K, Luo W, Sankaranarayana R. HTNet for micro-expression recognition. Neurocomputing. 2024;602(4):128196. doi:10.1016/j.neucom.2024.128196.
9. Yu K, Zhang Z, Hu C, Luo J. SOFP: capturing subtle facial dynamics with symmetric optical flow perception for micro-expression recognition. Pattern Recognit. 2026;176(12):113199. doi:10.1016/j.patcog.2026.113199.
10. Li J, Qian Y, Zhao L, Wang SJ. FED-PsyAU: privacy-preserving micro-expression recognition via psychological AU coordination and dynamic facial motion modeling. In: Proceedings of the 2025 IEEE/CVF International Conference on Computer Vision (ICCV); 2025 Oct 19–25; Honolulu, HI, USA. p. 14453–63.
11. Wang Z, Yang M, Jiao Q, Xu L, Han B, Li Y, et al. Two-level spatio-temporal feature fused two-stream network for micro-expression recognition. Sensors. 2024;24(5):1574. doi:10.3390/s24051574.
12. Ross ED, Gupta SS, Adnan AM, Holden TL, Havlicek J, Radhakrishnan S. Neurophysiology of spontaneous facial expressions: I. Motor control of the upper and lower face is behaviorally independent in adults. Cortex. 2016;76:28–42.
13. Iwasaki M, Noguchi Y. Hiding true emotions: micro-expressions in eyes retrospectively concealed by mouth movements. Sci Rep. 2016;6(1):22049.
14. Chaudhry R, Ravichandran A, Hager G, Vidal R. Histograms of oriented optical flow and Binet-Cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition; 2009 Jun 20–25; Miami, FL, USA. p. 1932–9.
15. Liu YJ, Zhang JK, Yan WJ, Wang SJ, Zhao G, Fu X. A main directional mean optical flow feature for spontaneous micro-expression recognition. IEEE Trans Affect Comput. 2016;7(4):299–310. doi:10.1109/taffc.2015.2485205.
16. Happy SL, Routray A. Fuzzy histogram of optical flow orientations for micro-expression recognition. IEEE Trans Affect Comput. 2019;10(3):394–406. doi:10.1109/taffc.2017.2723386.
17. Gan YS, Liong ST, Yau WC, Huang YC, Ken TL. OFF-ApexNet on micro-expression recognition system. Signal Process Image Commun. 2019;74(2):129–39. doi:10.1016/j.image.2019.02.005.
18. Wang Y, Huang Y, Liu C, Gu X, Yang D, Wang S, et al. Micro expression recognition via dual-stream spatiotemporal attention network. J Healthc Eng. 2021;2021:7799100. doi:10.1155/2021/7799100.
19. Zhai Z, Zhao J, Long C, Xu W, He S, Zhao H. Feature representation learning with adaptive displacement generation and transformer fusion for micro-expression recognition. In: Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17–24; Vancouver, BC, Canada. p. 22086–95.
20. Zhou L, Mao Q, Huang X, Zhang F, Zhang Z. Feature refinement: an expression-specific feature learning and fusion method for micro-expression recognition. Pattern Recognit. 2022;122:108275.
21. Khuong VTA, Nguyen LT, Man TBP, Nguyen MD, Le DD. DIANet: a phase-aware dual-stream network for micro-expression recognition via dynamic images. arXiv:2510.12219. 2025.
22. Liu S, Mao X, Zhao S, Li P, Xu T, Chen E. MER-CLIP: AU-guided vision-language alignment for micro-expression recognition. IEEE Trans Affect Comput. 2025;16(4):3028–42. doi:10.1109/TAFFC.2025.3584918.

23. Wei J, Sun J, Lu G, Yan J, Zhang D. Multi-information hierarchical fusion transformer with local alignment and global correlation for micro-expression recognition. In: Proceedings of the 33rd ACM International Conference on Multimedia; 2025 Oct 27–31; Dublin, Ireland. p. 5873–82.
24. Fan X, Chen X, Jiang M, Zhao Y, Liu Y, Yang J. SelfME: self-supervised motion learning for micro-expression recognition. In: Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17–24; Vancouver, BC, Canada. p. 13834–43.
25. Nguyen XB, Duong CN, Li X, Liu H, Nguyen TV, Luu K. Micron-BERT: BERT-based facial micro-expression recognition. In: Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17–24; Vancouver, BC, Canada. p. 1482–92.
26. Zhou H, Huang S, Li J, Wang SJ. Dual-ATME: dual-branch attention network for micro-expression recognition. Entropy. 2023;25(3):460.
27. Xie Z, Zhao C. Dual-branch cross-attention network for micro-expression recognition with transformer variants. Electronics. 2024;13(2):461. doi:10.3390/electronics13020461.
28. Zhang L, Zhang Y, Sun X, Tang W, Wang X, Li Z. Micro-expression recognition based on direct learning of graph structure. Neurocomputing. 2025;619(11):129135. doi:10.1016/j.neucom.2024.129135.
29. Wang F, Li J, Qi C, Wang L, Wang P. A multi-modal multi-scale network based on transformer for micro-expression recognition. J Vis Commun Image Represent. 2025;111(4):104537. doi:10.1016/j.jvcir.2025.104537.
30. Ma B, Wang L, Wang Q, Wang H, Li R, Xu L, et al. Entire-detail motion dual-branch network for micro-expression recognition. Pattern Recognit Lett. 2025;189:166–74. doi:10.1016/j.patrec.2025.01.021.
31. He J, Xiao Y, Zhang H, Cai J, Cai L, Liu R. Micro_NesT: multi-scale attention enhanced micro-expression recognition framework. Expert Syst Appl. 2025;290(1):128372. doi:10.1016/j.eswa.2025.128372.
32. Zach C, Pock T, Bischof H. A duality based approach for realtime TV-L1 optical flow. In: Pattern recognition. Berlin/Heidelberg, Germany: Springer; 2007. p. 214–23.
33. Liong ST, See J, Phan RCW, Ngo AL, Oh YH, Wong K. Subtle expression recognition using optical strain weighted features. In: Computer Vision–ACCV 2014 Workshops. Cham, Switzerland: Springer; 2015. p. 644–57.
34. Kazemi V, Sullivan J. One millisecond face alignment with an ensemble of regression trees. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition; 2014 Jun 23–28; Columbus, OH, USA. p. 1867–74.
35. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: transformers for image recognition at scale. In: Proceedings of the International Conference on Learning Representations; 2021 May 3–7; Virtual.
36. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Adv Neural Inf Process Syst. 2017;30:5998–6008. doi:10.65215/r5bs2d54.
37. Ba JL, Kiros JR, Hinton GE. Layer normalization. arXiv:1607.06450. 2016.
38. Li X, Pfister T, Huang X, Zhao G, Pietikäinen M. A spontaneous micro-expression database: inducement, collection and baseline. In: Proceedings of the 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG); 2013 Apr 22–26; Shanghai, China. p. 1–6.
39. Yan WJ, Li X, Wang SJ, Zhao G, Liu YJ, Chen YH, et al. CASME II: an improved spontaneous micro-expression database and the baseline evaluation. PLoS One. 2014;9(1):e86041.
40. Davison AK, Lansley C, Costen N, Tan K, Yap MH. SAMM: a spontaneous micro-facial movement dataset. IEEE Trans Affect Comput. 2018;9(1):116–29.
41. See J, Yap MH, Li J, Hong X, Wang SJ. MEGC 2019—the second facial micro-expressions grand challenge. In: Proceedings of the 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019); 2019 May 14–18; Lille, France. p. 1–5.
42. Yap MH, See J, Hong X, Wang SJ. Facial micro-expressions grand challenge 2018 summary. In: Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018); 2018 May 15–19; Xi'an, China. p. 675–8.

43. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: an imperative style, high-performance deep learning library. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems; 2019 Dec 8–14; Vancouver, BC, Canada. p. 8024–35.
44. Loshchilov I, Hutter F. Decoupled weight decay regularization. In: Proceedings of the International Conference on Learning Representations; 2019 May 6–9; New Orleans, LA, USA.
45. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Adv Neural Inf Process Syst. 2012;25(6):1097–105. doi:10.1145/3065386.
46. Wang F, Li J, Qi C, Wang L, Wang P. JGULF: joint global and unilateral local feature network for micro-expression recognition. Image Vis Comput. 2024;147:105091.
47. Zhang M, Yang W, Wang L, Wu Z, Chen D. HFA-Net: hierarchical feature aggregation network for micro-expression recognition. Complex Intell Syst. 2025;11(3):169.
48. Mao X, Liu S, Zhao S, Xu T, Chen E. MERba: multi-receptive field MambaVision for micro-expression recognition. arXiv:2506.14468. 2025.
49. Zhang Y, Lin W, Zhang Y, Xu J, Xu Y. Leveraging vision transformers and entropy-based attention for accurate micro-expression recognition. Sci Rep. 2025;15(1):13711. doi:10.1038/s41598-025-98610-y.
50. Zhang B, Wang X, Wang C, He G. Dynamic stereotype theory induced micro-expression recognition with oriented deformation. In: Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2025 Jun 10–17; Nashville, TN, USA. p. 10701–11.
51. Jiang H, Lyu J, Lan X, Xue J. Continuous action unit intensity modeling for micro-expression recognition. In: Proceedings of the 2025 IEEE International Conference on Image Processing (ICIP); 2025 Sep 14–17; Anchorage, AK, USA. p. 941–6.
52. Verma M, Vipparthi SK, Murala S, Abdel-Mottaleb M. ME-NAS: a micro expression feature adaptive neural architecture search. ACM Trans Intell Syst Technol. 2026;17(2):1–19.
53. Liu J, Shi H, Wang Y, Zheng W. FPS: frequency prompt synchronization for micro-expression recognition. Pattern Recognit. 2026;179:113572.
54. Shou Z, Huang L, Yuan X, Yu Y, Xu X, Wu Z. DSBICNet: dynamic-static bidirectional interaction collaborative network for micro-expression recognition. Multimed Syst. 2026;32(2):122.
55. Jie Z, Wei J, Feng Q, Wang S. LKDTNet: large kernel deconstruction three-dimensional network for micro-expression recognition. Signal Process Image Commun. 2026;143(5):117511. doi:10.1016/j.image.2026.117511.
56. Gan YS, Liu KH, Liong GB, Liong ST. Micro-expression recognition in wild video environments: latent feature-based ANN (LFANN) from 3D reconstructed faces. Neurocomputing. 2025;625:129480.
57. Wang H, Wang L, Xu L, Li Y. DBDE-Net: dual-branch detail-enhanced network for micro-expression recognition. Neurocomputing. 2026;679:133167.
58. Shuai T, Beng S, Khalid FB, Rahmat RWBOK. Advances in facial micro-expression detection and recognition: a comprehensive review. Information. 2025;16(10):876. doi:10.3390/info16100876.
59. Ahmad A, Li Z, Iqbal S, Aurangzeb M, Tariq I, Flah A, et al. A comprehensive bibliometric survey of micro-expression recognition system based on deep learning. Heliyon. 2024;10(5):e27392. doi:10.1016/j.heliyon.2024.e27392.