Regulatory Focus Theory Induced Micro-Expression Analysis with Structured Representation Learning

Anonymous Author(s)
Submission Id: 6431

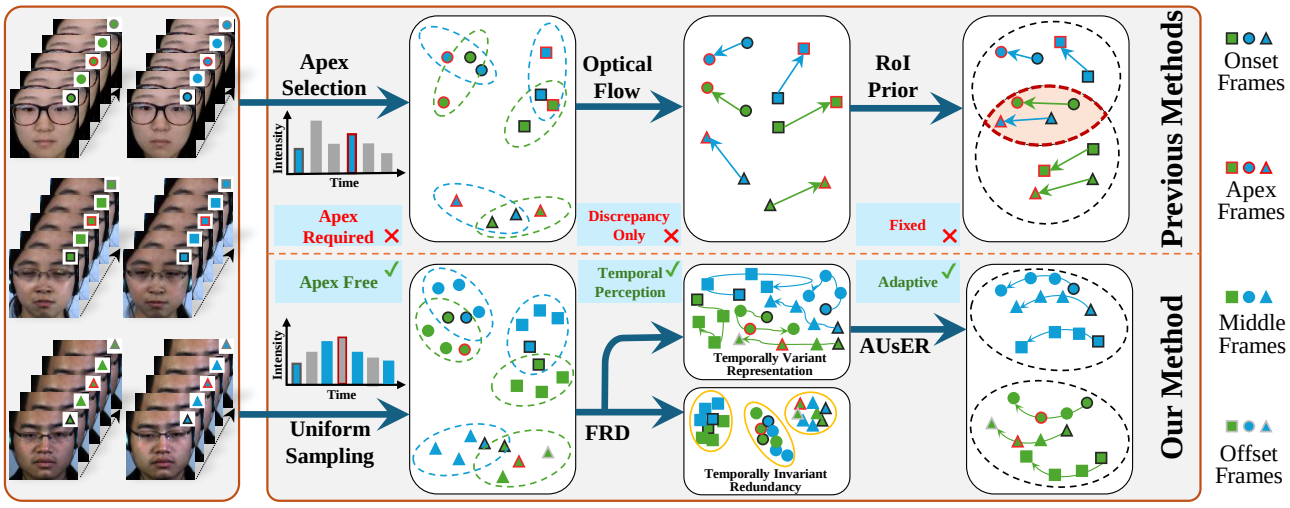


Figure 1: Comparison of apex-based methods and our apex-free framework: blue and green dashed ellipses denote MEs, yellow solid ellipses mark redundant information, and black dashed ellipses indicate similar patterns. The dark red dashed ellipse highlights that prior methods, lacking temporal perception and complete representations, often produce ambiguous groupings. Notably, our method eliminates apex reliance by uniformly sampling sequences, with only the onset frame guaranteed.

Abstract

Micro-expression analysis (MEA) is crucial for detecting subtle emotional cues, with applications in lie detection and psychological assessment. Existing methods struggle with three main challenges: 1) Noise sensitivity arising from the inherent subtlety of micro-expressions. 2) Reliance on fixed priors and apex annotations. 3) Information redundancy, with static features often dominating over dynamic emotional cues. To address these challenges, we propose Ac4AU, a framework inspired by Regulatory Focus Theory (RFT) that utilizes structured representation learning to decompose dynamic emotional patterns from redundant features. Specifically, AC4AU first leverages a face recognition backbone to extract robust yet redundant static representations. Secondly, a Frequency-aware Redundancy Decomposer (FRD) is introduced to eliminate the Direct Current component and retain the dynamic and process-sensitive features. Finally, a dynamic expert allocation mechanism, embodied by the AU-specific Expert Router (AUsER), is adopted to learn localized facial motion patterns and capture long-term relationships, enabling AU-targeted supervision and enhancing generalization across diverse datasets. Rigorous experiments demonstrate that the apex-free AC4AU achieves performance comparable to state-of-the-art apex-dependent methods. Additionally, we conduct a statistical analysis that provides insights into the AU dependencies. Code will be made available upon request.

CCS Concepts

• **Computing methodologies** → **Activity recognition and understanding**; *Motion capture*; Cross-validation.

Keywords

Micro Expression Analysis, Action Units Detection, Affective Computing, Computer Vision, Mixture of Experts

1 Introduction

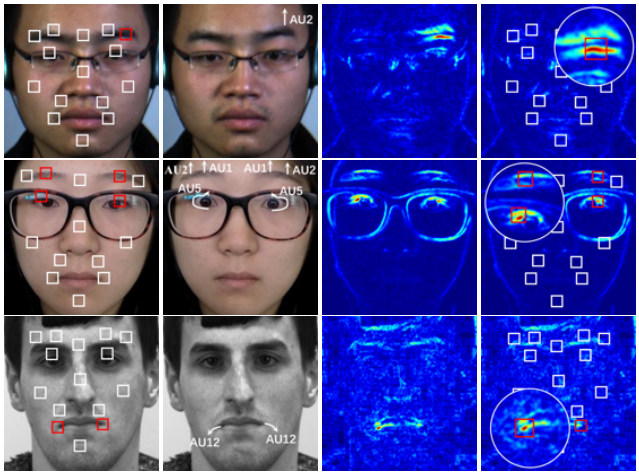
The immense potential of micro-expressions (MEs) in lie detection, public safety and other specific scenarios has attracted widespread attention [27, 31, 46], particularly in the area of micro-expression analysis (MEA), which aims to reveal inner activities and concealed emotions [23]. However, the subtle and rapid movements [8, 45] of MEs, which are nearly invisible to the naked eye, pose unique and significant challenges for the research community [25, 41, 53]. To overcome these inherent challenges, most previous methods leverage deep learning and computer vision techniques, achieving outstanding performance [1, 21, 51]. Additionally, some methods have incorporated other advanced techniques such as multimodal ensemble learning [49], progressive learning [39], causally uncovering [33] and prototypical contrastive learning [11]. Maintaining a focus on various techniques, prominent experts have noticed that some studies exhibit "somewhat questionable" issues, undermining the reliability of traditional ME recognition (MER) [2]. Therefore,

they introduce CD6ME [35], a new benchmark treats action unit (AU) [9] detection (AUD) [24] as a distinct MEA task.

Unlike MER, which classifies MEs into emotional categories, AUD focuses on the most granular visual units of MEs. It is worth noting, though easily understandable, that while AUD is a more fine-grained task, the inherent challenges it faces are not only similar to those in MER, but are even more formidable. **1) The issue of representation ambiguity** caused by the unclear distinction between MEs and unrelated head movements or other noises. Therefore, facial alignment is commonly adopted for explicit denoising [20, 30]. **2) High information redundancy**, which also stems from the intrinsic properties of MEs. Frame-level representations often contain a large amount of irrelevant information, resulting in high redundancy that impairs the efficiency of ME representation. To address this issue, most methods utilize optical flow between the onset (first) and apex (peak intensity) frames to construct the representation [12, 26], leading to reliance on apex annotations. Another common strategy is to leverage prior knowledge [15, 17] to define regions of interest (RoIs), from which features are selectively extracted, , but this often leads to inadequate spatial representations. **3) The sample imbalance problem** [13], which arises from the uneven distribution of samples across datasets. This imbalance often leads to biased training and can significantly impact the generalization. Some methods leverage self-supervised learning [5, 10] and additional samples [25, 41] to reduce the reliance on labeled data and enhance model generalization across imbalanced classes.

Despite the progress achieved by previous methods, MEA still faces unresolved challenges—either stemming from its inherent characteristics that make effective representation difficult, or arising from new issues introduced by recently adopted techniques. **1) Representation ambiguity from imperfect alignment.** Explicit facial alignment may fail to accurately align subtle facial movements but instead introduce additional noise; **2) Temporal under-representation in pair-wise motion modeling.** As Figure 1 highlights, optical flow computed between onset and apex frames overlooks the dynamics of intermediate frames [35]. Meanwhile, as further discussed in Appendix A, the substantial variation in temporal lengths of MEs exacerbates the instability of optical flow representations computed from frame pairs; **3) Inadequate spatial representation from fixed priors based RoIs.** Dividing the face into predefined RoIs may overlook cross-region interactions, restricting the ability of capturing subtle spatial dependencies critical for AU activation patterns. As shown in Figure 2, inadequate representation is evident since such fixed regions may fail to capture subtle or off-region patterns of MEs; **4) Generalization challenges from intensified imbalance and variable duration.** As discussed in Appendix B and Figure 4, AUD faces pronounced sample imbalance across AUs and substantial variation in temporal durations, undermining the generalization ability.

To address the aforementioned challenges, we propose AC4AU, a novel structured representation learning network for apex-free action unit detection (AUD). Inspired by the Regulatory Focus Theory [3, 14], which posits that emotional experiences are guided by both promotion and prevention motivations, we emphasize the dynamic nature of emotional processes. RFT suggests that emotions evolve continuously over time as individuals adapt their behavior to achieve desired goals or avoid undesirable outcomes. This



(a) Onset (b) Apex (c) Optical Flow (d) RoI Selection

Figure 2: Illustration of the typical steps in previous apex-based methods. Specifically, (a): the onset frame with predefined RoIs; (b): the apex frame annotated with AU activations; (c): optical flow between onset and apex frames, which adopted by most previous MEA methods; (d): local representations extracted within the RoIs, primarily those marked in red, while the majority of the white-marked RoIs fail to capture meaningful representations.

theoretical framework supports the need for dynamic representation learning to effectively capture the temporal progression of MEs, ensuring that both emotion-related dynamics and temporal context are accurately modeled. To realize the dynamic structured representation learning, we introduce a series of components in AC4AU, each designed to address specific challenges associated with micro-expression analysis. Firstly, to enhance robustness and mitigate the additional noise introduced by explicit facial alignment, we leverage the noise-resilience demonstrated by face recognition models [7] and incorporate a pre-trained face recognition model as the backbone of AC4AU. Secondly, in line with the emphasis of RFT on the continuous nature of emotional dynamics, we argue that using only onset and apex frames disrupts this continuity. To preserve temporal progression, we uniformly sample the entire sequence to a fixed length, as shown in Figure 1. To further suppress redundancy, we introduce the Frequency-aware Redundancy Decomposer (FRD), which removes the Direct Current (DC) component and retains Alternating Current (AC) components, capturing dynamic emotional variations. Thirdly, to overcome the limitations of prior-based feature selection—often relying on empirical assumptions and failing to guarantee complete spatial representation—we introduce the AU-specific Expert Router (AUsER), which is built upon the Mixture of Experts (MoE) framework. AUsER facilitates dynamic, task-aware expert routing by adaptive representation selection, enabling experts to focus on AU-specific features. This mechanism enables each expert to focus on learning localized facial motion patterns associated with a particular AU, resulting in more discriminative and accurate representation learning. Finally, to address the sample imbalance in AUD, we adopt Focal Loss to

down-weight easy negatives and emphasize informative hard examples. Combined with AUsER, which assigns experts to specific AUs, this supervision enables AU-targeted gradient updates. Unlike joint optimization over all AUs, which is often affected by co-occurrence or mutual exclusivity, our approach provides clean supervision to each expert, mitigating optimization conflicts and enhancing generalization across imbalanced distributions.

To summarize, the primary contributions are as follows:

- We propose AC4AU, an apex-free network for AUD that addresses data leakage concerns raised in CD6ME and eliminates the need for manually annotated apex frames. AC4AU integrates frequency-based representation decomposition with AU-specific expert routing, enabling dynamic and noise-resilient ME representations.
- To reduce redundancy in ME representations, we introduce the Frequency-aware Redundancy Decomposer (FRD), which removes the Direct Current (DC) component and retains dynamic emotional variations via the Alternating Current (AC) component. Inspired by Regulatory Focus Theory (RFT), FRD aligns with the dynamic nature of emotional expressions, ensuring the preservation of process-sensitive features crucial for AUD.
- We address the limitations of prior-based feature selection with the introduction of the AU-specific Expert Router (AUsER), combined with Focal Loss to mitigate sample imbalance. AUsER, based on the Mixture of Experts (MoE) framework, learns localized motion patterns for each AU, while Focal Loss provides AU-targeted supervision to improve both accuracy and generalization.
- Extensive experiments validate the performance of our method, demonstrating fair comparison and robust cross-dataset generalization. Additionally, a statistical analysis of AU co-occurrence and mutual exclusivity provides insights into AU dependencies, further validating the effectiveness of AU-specific structured representation learning.

2 Related Work

Due to the limited attention AUD has received in the past, most existing methods have focused on traditional MER, aiming to classify expressions into predefined categories. These approaches, while effective in some cases, often overlook the finer nuances of ME features critical for comprehensive analysis or effective classification.

2.1 Handcrafted Methods

Handcrafted feature-based methods have been widely explored in MER due to their interpretability and computational efficiency. MDMO (Main Directional Mean Optical Flow) [28] calculates the mean optical flow in the principal directions of facial regions, effectively capturing subtle movements in MEs. However, its performance can be limited when dealing with complex and diverse motion patterns. LBP-TOP (Local Binary Patterns on Three Orthogonal Planes) [54] extracts spatiotemporal texture features by analyzing local patterns across spatial and temporal dimensions, demonstrating its robustness in dynamic texture recognition tasks. Despite its success, it may struggle with the subtlety of MEs. Bi-WOOF (Bi-Weighted Oriented Optical Flow) [27] enhances sensitivity to subtle changes by leveraging optical flow direction, magnitude and strain, combined with a local-global weighted strategy. Nevertheless, the handcrafted nature of these methods often limits their ability to generalize effectively across diverse datasets.

2.2 Deep Learning-based Methods

Deep learning has demonstrated remarkable performance across various tasks in recent years. The micro-expression community has quickly adopted a variety of deep learning-based methods and training paradigms to address the unique challenges posed by MEs.

Convolutional neural networks (CNNs) have played a pivotal role in early deep learning advancements in MER [50]. OfApexNet [12] leverages the apex frame in ME sequences to emphasize key temporal moments, significantly enhancing robustness in capturing subtle motions. Similarly, STSTNet [26] utilizes a shallow triple-stream 3D CNN to process multiple types of information, effectively balancing computational efficiency with feature representation. While these approaches have proven effective, they face challenges in fully exploiting the dynamic nature of MEs due to their limited temporal modeling capabilities. To address these limitations, more recent methods have explored attention mechanisms and graph-based techniques, which offer enhanced flexibility in modeling long-range dependencies and complex relationships.

Attention mechanisms and graph-based methods have further advanced the field by modeling relationships between local facial landmarks [15, 17]. These approaches effectively capture region-specific dynamics but rely on well-annotated landmarks. To complement these spatial advancements, adversarial training [42] has been introduced to decouple identity and ME, thereby improving robustness in cross-dataset scenarios. While attention mechanisms and graph-based methods have been widely explored, **transformer-based** methods have shown great promise in handling long-term dependencies and capturing subtle motion. Micron-BERT [29] utilizes a transformer-based encoder for sequence modeling, while feature representation learning with adaptive displacement generation and transformer fusion [48] enhances temporal modeling by combining local and global temporal patterns. Additionally, temporal-informative adapters in VideoMAE [47] improve multi-scale feature fusion, further demonstrating the potential of transformers in MER.

Self-supervised learning has been instrumental in addressing the challenges posed by limited labeled data. SelfMe [10] leverages self-supervised motion pretraining to improve cross-dataset generalization, while interpretable self-supervised facial micro-expression learning [5] extends these principles to predict cognitive states and neurological disorders. Additionally, MMNet [18] emphasizes muscle motion-guided features, providing a unique approach for modeling subtle muscle movements specific to micro-expressions. Other innovative techniques include CMNet [40], amplifying subtle features through contrastive magnification and three-stream graph attention networks [16], employing dynamic patch selection for expression-specific feature extraction. Methods addressing data imbalance [13] and noisy labels [36] tackle practical challenges, while LED [37] and SCA [24] focus on spatiotemporal feature refinement. The creation of high-resolution 3D dynamic facial expression datasets [52] has significantly enriched the field by

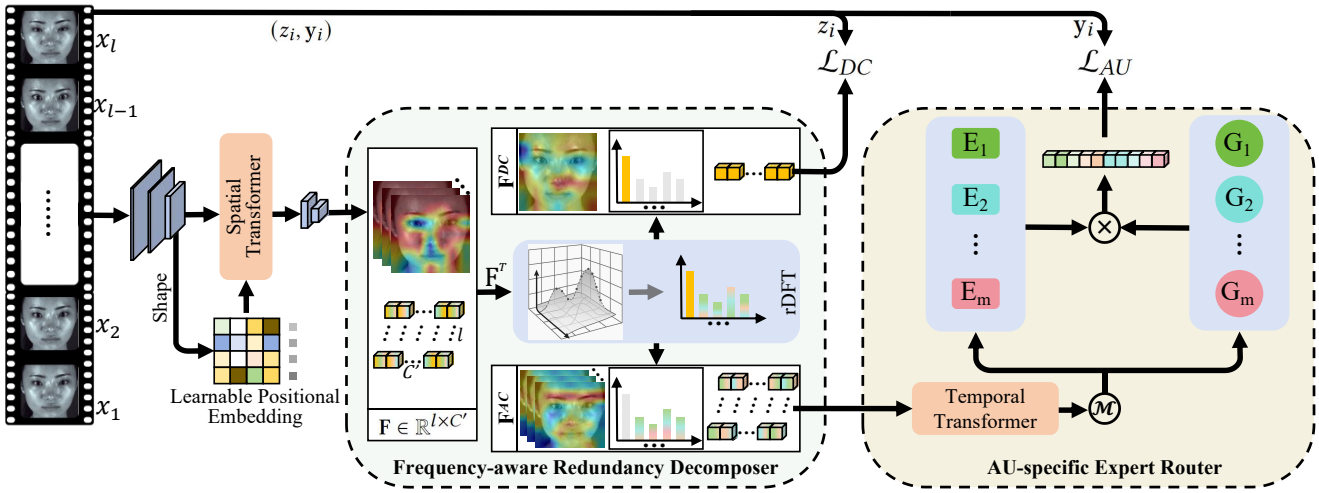


Figure 3: Overview of the proposed AC4AU for apex-free AU activation detection from micro-expression sequences. Given a uniform sampled sequence $\{x_1, \ldots, x_l\}$, frame-level features are first extracted using a MobileFaceNet-based backbone, into which a spatial transformer with learnable positional embeddings is integrated to enhance noise-robust spatial representation. The resulting sequence $\mathbf{F} \in \mathbb{R}^{l \times C'}$ is then processed by the Frequency-aware Redundancy Decomposer (FRD), which separates temporally invariant redundancy $\mathbf{F}^{DC}$ and emotion-relevant dynamics $\mathbf{F}^{AC}$ via rDFT. The resulting dynamics is passed through a temporal transformer and routed by the AU-specific Expert Router (AUsER) for final prediction. The colored cubes represent different representations: yellow denotes temporally invariant redundancy, while others indicate emotional dynamics.

providing high-quality data. Furthermore, data-centric approaches such as DC-check [32] provide guidelines to ensure reliable machine learning systems, emphasizing the importance of robust.

In summary, advancements from CNN-based methods to transformers, graph models and self-supervised learning highlight the rapid evolution of MER. However, existing MER methods suffer from potential data leakage issues, undermining fair comparisons. Moreover, AUD remains in its infancy, with considerable room for development toward more comprehensive and equitable analysis. Notably, benchmark datasets play a critical role in advancing the field. The CD6ME [35] benchmark integrates 6 ME datasets [2, 6, 19, 22, 43, 44] into a unified standard for fair evaluation, addressing label inconsistency and enabling rigorous cross-dataset validation.

3 Methodology

Action unit detection from micro-expressions presents unique challenges due to subtle motion and short duration. To address these issues, we design a unified framework named AC4AU, as illustrated Figure 3. The proposed AC4AU consists of three main components: a noise-robust spatial representation module as backbone, a frequency-aware redundancy decomposer, a temporal modeling and AU-specific expert routing module. In addition, a tailored loss function is introduced to jointly supervise AU prediction and redundancy representation decomposition. Together, these components enable apex-free, adaptive, and generalizable AU detection.

3.1 Problem Formulation

We formulate Action Unit Detection (AUD) from micro-expression video sequences as a multi-label binary classification task. Given a dataset $\mathcal{V} = \{v_i^{var}\}_{i=1}^{N}$ consisting of N variable-length video sequences, we uniformly sample each v_i^{var} to a fixed length to ensure consistent temporal input v_i. Each resampled sequence v_i is composed of l consecutive RGB frames $\{x_j\}_{j=1}^{l}$, where $x_j \in \mathbb{R}^{3\times112\times112}$. Besides, v_i is annotated with a binary vector $\mathbf{y}_i = [y_1, y_2, \ldots, y_{12}] \in \{0,1\}^{12}$, where $y_k = 1$ indicates the activation of the k-th Action Unit (AU). The objective of AUD is to learn a mapping function $f : \mathbb{R}^{l\times3\times112\times112} \rightarrow [0,1]^{12}$ that predicts the AU activation vector $\hat{\mathbf{y}}_i = f(v_i)$ for each input sequence v_i. This task presents substantial challenges, including severe noise interference arising from high information redundancy and task-irrelevant factors, inadequate spatial representations of subtle facial motions and significant label imbalance of AU distribution.

To further guide the representation learning process, we assume that the redundant information in MEs are subject-invariant and consistent across time and samples. Therefore, each sequence v_i is additionally annotated with an subject label $z_i \in \{1, 2, \ldots, S\}$, where S denotes the total number of unique subjects in the training datasets. This label serves as supervision for decomposing redundant components in frequency-aware redundancy decomposer. The final output is a vector $\hat{\mathbf{y}}_i = [\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_{12}] \in \{0,1\}^{12}$, where $\hat{y}_k$ denotes the predicted probability that the k-th AU is activated. Following standard binary classification practices, a threshold of 0.5 is applied to determine AU activation.

3.2 Noise-Robust Spatial Representation

To extract robust and expressive spatial features from MEs, we adopt a pre-trained face recognition backbone with proven robust to head movements, lighting variations, and other noise. Importantly, instead of relying on explicit facial alignment, which may

introduce additional noise and artifacts, we enhance the backbone by structurally integrating spatial dependency into the backbone network. Specifically, we modify the standard MobileFaceNet [4] architecture by introducing a spatial transformer [38] module that captures long-range dependencies between facial regions.

Let $x \in \mathbb{R}^{B\times l\times 3\times 112\times 112}$ denote a batch of B resampled MEs. We first merge the batch and temporal dimensions and extract spatial features via a series of convolutional layers. The resulting features are reshaped into a sequence of tokens $\mathbf{X} \in \mathbb{R}^{(B\times l)\times(H\times W)\times C}$, where each of the $H \times W$ tokens represents a specific facial region. To encode spatial location, we add a learnable positional embedding $\mathbf{E} \in \mathbb{R}^{1\times(H\times W)\times C}$ and obtain the enriched input as $\mathbf{X} \leftarrow \mathbf{X} + \mathbf{E}$. The spatial transformer encoder, equipped with 4 attention heads, updates $\mathbf{X}$ via multi-head self-attention. The output is further refined through a lightweight residual mapping, consisting of three convolutional layers interleaved with batch normalization and a PReLU activation after the first normalization. This process yields the spatially enhanced frame-level representation $\mathbf{F} \in \mathbb{R}^{(B\times l)\times C'}$.

3.3 Frequency-aware Redundancy Decomposer

To suppress redundancy information and enhance the discriminability of dynamic facial cues, we introduce the Frequency-aware Redundancy Decomposer (FRD), as shown in Figure 3. This performs a frequency-domain decomposition of frame-level spatial representations to separate temporally invariant (static) and variant (dynamic) components. The underlying assumption is that redundant information irrelevant to AU activation are mostly subject-invariant and reside in the temporally stationary (i.e., DC) components of micro-expression sequences. Let $\mathbf{F} \in \mathbb{R}^{B\times l\times C'}$ denote the frame-level spatial features extracted in the previous stage, where B is the batch size, l is the number of uniformly sampled frames, and C' is the feature dimension per frame. For clarity, we omit the batch dimension and analyze a single sequence with $\mathbf{F} \in \mathbb{R}^{l\times C'}$.

We perform a real-valued Discrete Fourier Transform (rDFT) along the temporal dimension. For each channel $c \in \{1, \ldots, C'\}$, the temporal signal $F_{t,c}|_{t=0}^{l-1}$ is transformed into its frequency-domain representation $\widehat{F}k, c$:

$$\begin{aligned}\widehat{F}_{k,c} &= \sum_{t=0}^{l-1} F_{t,c} \cdot e^{-2\pi ikt/l}, \quad k = 0, 1, \ldots, \left\lfloor \frac{l}{2} \right\rfloor, \\ &= \sum_{t=0}^{l-1} F_{t,c} \cdot \left[\cos\left(\frac{2\pi kt}{l}\right) - i \cdot \sin\left(\frac{2\pi kt}{l}\right)\right],\end{aligned} \tag{1}$$

where t indexes the temporal position, k denotes the frequency index, and $i = \sqrt{-1}$ is the imaginary unit. Due to the conjugate symmetry of the rDFT for real-valued signals, all frequency information is retained by preserving only the first $\left\lfloor \frac{l}{2} \right\rfloor + 1$ coefficients.

Considering the earlier assumption that redundant information is temporally invariant and subject-consistent, we define the DC component as the zero-frequency term:

$$\boldsymbol{F}^{\mathrm{DC}} = \Re\left(\widehat{\mathbf{F}}_0\right) \in \mathbb{R}^{C'}, \tag{2}$$

which captures temporally invariant representations and is later used to supervise subject-level consistency.

To extract dynamic variations, we retain only the non-zero frequency components and apply inverse rDFT to represent the temporally resolved motion patterns:

$$\mathbf{F}_t^{\mathrm{AC}} = \Re\left(\sum_{k=1}^{\lfloor l/2 \rfloor} \widehat{F}_{k,c} \cdot e^{2\pi ikt/l}\right), \quad t = 0, 1, \ldots, l-1, \tag{3}$$

yielding the AC-enhanced features $\mathbf{F}^{\mathrm{AC}} \in \mathbb{R}^{l\times C'}$.

With this decomposition, the original temporal features are separated into a static component $\mathbf{F}^{\mathrm{DC}}$ representing subject-invariant redundancy, and a sequence of dynamic components $\mathbf{F}_t^{\mathrm{AC}}|_{t=0}^{l-1}$ encoding ME-relevant temporal variations.

3.4 Action Unit Specific Expert Router

To capture the heterogeneous nature of facial action units, we introduce an Action Unit Specific Expert Router (AUsER), which adaptively allocates specialized experts to different AUs. Unlike prior works that rely on predefined Regions of Interest (RoIs) for each AU, AUsER learns to dynamically select AU-relevant features from the shared representation space. This is inspired by the observation that different AUs exhibit distinct spatial and motion patterns, as discussed in Section 4.4. AUsER is instantiated as a Mixture-of-Experts (MoE) architecture composed of twelve parallel experts $\{\mathbf{E}_m(\cdot)\}_{m=1}^{12}$, each dedicated to one AU. Given the dynamic representation $\mathbf{F}^{\mathrm{AC}} \in \mathbb{R}^{l\times C'}$, a temporal transformer encoder $\mathcal{T}$ is applied to model long-range dependencies, and average pooling $\mathcal{M}$ is used to aggregate sequence-level semantics:

$$\mathbf{f} = \mathcal{M}\left(\mathcal{T}\left(\mathbf{F}^{\mathrm{AC}}\right)\right) \in \mathbb{R}^{C'}. \tag{4}$$

This global feature $\mathbf{f}$ is routed through a Mixture-of-Experts framework comprising 12 AU-specific experts. Each AU prediction is computed by a weighted combination of expert outputs:

$$\hat{y}_k = \sum_{m=1}^{12} \alpha_{k,m} \cdot E_m(\mathbf{f}), \tag{5}$$

where $\alpha_{k,m} = \mathrm{softmax}\left(G_k(\mathbf{f})\right)_m$ denotes the gating weight assigned to the m-th expert by the gating network G_k of k-th AU. Each gating network is implemented as a single-layer perceptron:

$$G_k(\mathbf{f}) = \mathbf{W}_k^{\mathrm{gate}}\,\mathbf{f} + \mathbf{b}_k^{\mathrm{gate}}, \quad k = 1, 2, \ldots, 12, \tag{6}$$

where $\mathbf{W}_k^{\mathrm{gate}} \in \mathbb{R}^{12\times C'}$ and $\mathbf{b}_k^{\mathrm{gate}} \in \mathbb{R}^{12}$ are learnable parameters, shared across training instances but specific to each AU. The final AU prediction vector is given by $\hat{\mathbf{y}} = [\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_{12}] \in [0, 1]^{12}$, where each $\hat{y}_k$ is computed via a weighted combination of AU-specific experts as shown above.

This design provides a flexible yet efficient mechanism to handle AU heterogeneity. The gating network adaptively modulates expert contributions based on input semantics, enabling each AU to select relevant knowledge from the expert pool. By decoupling feature specialization from region priors, AUsER achieves more robust modeling under subtle motion patterns. Moreover, its cooperative training with focal loss facilitates specialization of experts and routing strategies according to AU difficulty and imbalance, leading to enhanced generalization and fine-grained AU discrimination.

3.5 Loss Functions

To supervise both the learning of subject-invariant static representations and AU-specific dynamic features, we define a dual-branch loss function tailored to our model design.

As introduced in Section 3.3, the temporally invariant component $\mathbf{F}^{\mathrm{DC}} \in \mathbb{R}^{C'}$ extracted by FRD is used to encourage consistency among sequences belonging to the same subject. We apply an angular-margin-based classification loss to supervise $\mathbf{F}^{\mathrm{DC}}$:

$$\mathcal{L}_{DC} = -\log \frac{e^{32\cdot(\cos(\theta_{z_i}+0.5))}}{e^{32\cdot(\cos(\theta_{z_i}+0.5))} + \sum_{j\neq z_i} e^{32\cdot\cos(\theta_j)}}, \tag{7}$$

where θ_j is the angle between the normalized $\mathbf{F}^{DC}$ and the j-th subject prototype $\mathbf{w}_j$, a learnable unit vector representing the j-th subject in the embedding space. Since the number of unique subjects varies across folds under the leave-one-dataset-out (LODO) protocol, this supervision is applied only during training. Moreover, to address the severe imbalance across AU categories, we apply a class-aware focal loss to the predicted AU probability vector $\hat{\mathbf{y}}_i$ and the binary AU activation label $\mathbf{y}_i$:

$$\begin{aligned}\mathcal{L}_{AU} = -\sum_{k=1}^{12} \Big[&(1-\alpha_k)(1-\hat{y}_{i,k})^2 y_{i,k}\log(\hat{y}_{i,k}) \\ &+\alpha_k \hat{y}_{i,k}^2 (1-y_{i,k})\log(1-\hat{y}_{i,k})\Big],\end{aligned} \tag{8}$$

where α_k denotes the prior probability of positive samples for the k-th AU, independently estimated from the positive-negative ratio of the k-th AU, without being influenced by other AUs. Empirically, using the positive sample ratio helps emphasize rare AUs, improving generalization in the presence of severe label imbalance.

The final loss combines both objectives:

$$\mathcal{L} = \gamma \cdot \mathcal{L}_{DC} + \mathcal{L}_{AU}, \tag{9}$$

where γ is a balancing coefficient that modulates the contribution of static representation supervision.

4 Experiments

4.1 Implementation Details

In our experiments, we use MediaPipe [1] to detect facial landmarks and crop face regions, which are then resized to $\mathbb{R}^{3\times112\times112}$. As shown in Figure 4, sequence lengths vary significantly, while shorter sequences tend to cluster around certain lengths. To avoid introducing interpolated frames that do not exist in the original videos, we uniformly sample all sequences to a fixed length of $l = 8$, resulting in an input dimension of $\mathbb{R}^{8\times3\times112\times112}$. For sequences shorter than l, we replicate the last frame to meet the required length. To ensure a fair comparison and avoid data leakage, we strictly adhered to the leave-one-dataset-out (LODO) protocol [35]. The preprocessing and evaluation framework were uniformly applied across all datasets. For data augmentation, we employed horizontal flipping, random rotating and color jittering to enhance model generalization. All experiments were conducted using a single NVIDIA L40 GPU.

[1] https://github.com/google/mediapipe

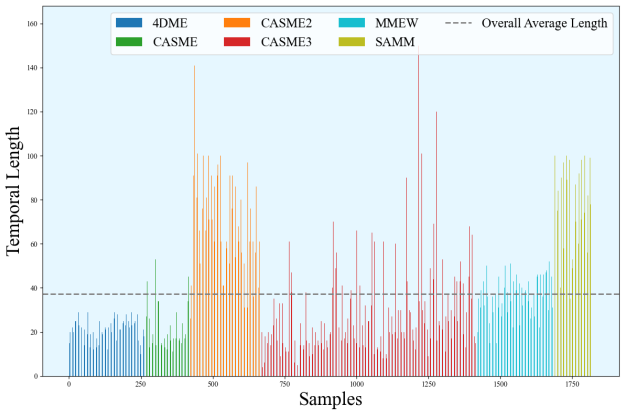


Figure 4: Illustration of temporal length distributions, a few excessively long samples are truncated for clarity.

4.2 Comparison with State-of-the-arts

To ensure a fair and consistent evaluation, we follow the unified protocol and experimental results provided by CD6ME, which relies on apex frame annotations for all baseline methods. In contrast, our method does not require apex labels and instead adopts uniform temporal sampling, enabling a fully apex-free setting across all datasets. Despite the stricter setting, AC4AU achieves comparable or superior performance, demonstrating strong generalization without relying on manually labeled apex frames.

As shown in Table 1, while several baselines perform well on individual AUs, none dominates across all categories, reflecting the inherent difficulty of AUD. Our method achieves the highest average F1 score and ranks first or second on 6 out of 12 AUs, showing consistent advantages. Notably, we observe significant gains on underrepresented AUs such as AU5, AU10, AU14, and AU15, suggesting the effectiveness of our approach in capturing subtle and sparse patterns. The detailed results of the 6-fold cross-validation under LODO are available in Appendix E, demonstrating the robustness and consistency of AC4AU across multiple folds. Importantly, all our experimental settings are statistically grounded and deliberately designed to avoid tuning tricks. The temporal length l used in our model is chosen based on dataset-wide analysis: overly long l force repetition of the last frame due to insufficient ME duration, while overly short ones compromise temporal representation. Furthermore, instead of hand-tuning the focal loss parameter α, we compute it once globally from all datasets and keep it fixed throughout all experiments to ensure fair and reproducible comparison.

4.3 Ablation Study

To validate the effectiveness and necessity of each proposed component and design choice—particularly in enhancing robustness and improving AU-discriminative representation—we conduct a series of ablation studies under the same evaluation protocol. Specifically, when ablating the focal loss, we replace it with standard binary cross-entropy (BCE) loss. For the FRD, we retain the full network

Table 1: Comparison with State-of-the-arts under the LODO protocol

Method	AU1	AU2	AU4	AU5	AU6	AU7	AU9	AU10	AU12	AU14	AU15	AU17	Average
LBP-TOP†	41.6	36.7	62.0	0.0	0.0	0.0	1.7	0.0	0.0	3.5	0.0	0.0	12.125
MDMO†	59.4	57.0	72.2	16.7	2.3	18.1	10.9	0.0	27.4	25.5	13.0	27.8	27.525
GA-ME†	71.1	67.6	85.7	1.1	0.0	26.2	7.8	2.1	12.8	15.4	0.0	34.7	27.042
Off-ApexNet†	74.9	70.2	86.3	13.5	3.3	44.5	**36.6**	18.3	32.0	37.7	19.0	38.2	39.542
STSTNet†	74.2	68.8	85.7	3.6	6.4	29.0	18.6	9.0	16.7	29.0	11.8	28.3	31.758
RCN-A†	74.8	71.0	85.3	4.7	0.0	24.1	15.9	0.0	21.1	23.7	0.0	24.0	28.717
NMER†	19.1	19.2	43.3	9.3	6.3	9.1	12.5	4.8	18.1	22.6	3.3	6.4	14.500
SSSNet†	74.6	**72.1**	**87.5**	13.6	5.3	**48.6**	20.3	19.3	36.2	40.8	22.7	**44.7**	40.475
ResNet10†	68.7	65.2	83.8	8.3	6.4	39.6	11.0	6.8	31.0	29.6	8.8	39.6	33.233
ResNet18†	**76.2**	**72.1**	87.1	12.6	5.5	38.2	8.6	18.4	**37.3**	33.7	12.4	**44.7**	37.233
ResNet34†	72.1	71.7	85.8	11.2	5.4	38.7	16.0	13.5	36.5	39.9	14.8	39.2	37.067
SCA†	42.3	42.8	56.2	14.8	1.2	23.4	13.4	3.0	21.8	37.1	4.0	22.9	23.575
LED†	52.7	45.7	63.7	7.9	0.7	19.3	13.6	8.5	26.5	36.7	33.0	31.7	28.333
RNet18(2+1)D†	54.2	49.4	72.7	26.9	5.2	20.4	11.7	12.4	23.3	25.5	9.7	41.9	29.442
AC4AU (Ours)	58.9	57.1	77.2	**44.8**	**6.8**	34.2	30.9	**25.0**	34.5	**47.5**	**36.7**	32.5	**40.508**

[1] Best results are in **bold**, second-best are underlined.
[2] Results† are sourced from the CD6ME benchmark [35].
[3] AU-wise scores are computed as the overall F1 scores across all six datasets.

Table 2: Ablation Studies

Pre-training	FRD	AUsER	Focal Loss	Overall F1
-	✓	✓	✓	22.270
✓	-	✓	✓	27.226
✓	✓	-	✓	29.647
✓	✓	✓	-	32.046
✓	✓	✓	✓	40.508

Table 3: Sensitivity Analysis of γ

γ	0.1	0.2	0.3	0.5	0.7
Overall F1	40.508	36.839	35.192	34.528	30.598

architecture but remove the decomposition operation. When evaluating the AUsER, we retain the temporal transformer for fairness and only remove the routing mechanism.

As shown in Table 2, all components contribute positively to the overall performance. Comparing row 4 and row 5 demonstrates that focal loss effectively addresses the sample imbalance issue, leading to a noticeable improvement. Meanwhile, the comparison between row 3 and row 5 highlights the benefit of AUsER, which enhances the discriminative ability by dynamically routing AU-specific features. These results jointly indicate that the combination of AUsER and focal loss is crucial for accurate and robust AUD. Besides, the first three rows show that each individual module—including the noise-robust face recognition backbone, FRD, and AUsER—offers consistent performance gains, while the integrated model with all components yields the best performance.

(a) Onset (b) Optical Flow (c) **F** (d) $\mathbf{F}^{AC}$

Figure 5: Comparative visualization of ME features extracted by optical flow and our method. (a) Onset frame with AU annotations. (b) Optical flow adopted by previous methods. (c) General facial representations retaining temporal redundancy. (d) AC component highlighting emotional dynamics.

4.4 Analysis and Discussion

To further evaluate the effectiveness and interpretability of AC4AU, we conduct in-depth analyses beyond performance comparisons. These analyses include a sensitivity analysis of γ, visualizations of discriminative and process-aware representations. In addition, we perform a statistical analysis of AU co-activation patterns, which is conducted independently of any model inference. This analysis offers complementary insights into the underlying structure and distribution of AU annotations across different datasets.

Sensitivity Analysis of γ. Table 3 presents an analysis of the sensitivity of the loss function to the parameter γ. The results indicate that $\gamma = 0.1$ consistently achieves the highest performance.

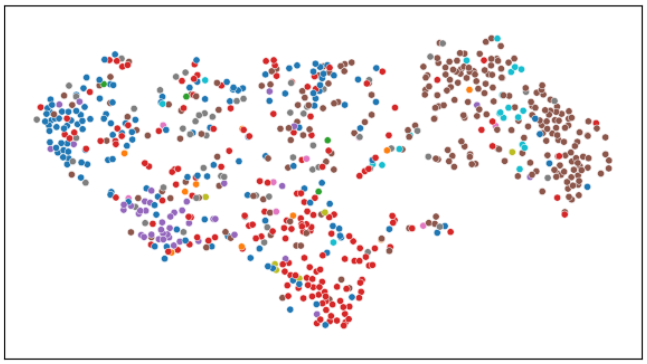


Figure 6: Visualization of dynamic emotional representations $\mathbf{F}^{AC}$, with colors indicating different AU activation patterns.

Visualization and Interpretability. As illustrated in Figure 5, previous methods rely on optical flow to capture motion dynamics, but their results are less clear and less focused compared to our method. Specifically, the optical flow visualization fails to highlight precise regions of interest, resulting in a less effective representation of the dynamic emotional cues. In contrast, The AC component $\mathbf{F}^{AC}$ we adopte not only removes redundant information effectively but also adaptively extracts features from regions surrounding activated AUs, with the backbone feature $\mathbf{F}$ captures a complete facial representation but contains substantial spatial redundancy. To further examine the representational structure of our model, we visualize the decomposed features $\mathbf{F}^{AC}$ and $\mathbf{F}^{DC}$ using t-SNE [34], as shown in Figure 6. For visualization purposes, the 12-dimensional AU activation vectors are treated as binary codes and interpreted as categorical AU activation patterns. While several cluster centers are clearly observable in the figure—suggesting good separability—this contradicts the corresponding quantitative performance on CASME3 reported in Appendix E. Upon closer analysis, we attribute this inconsistency to the high sample volume in CASME3 and low intrinsic separability of AU activation patterns, which leads to significant overlap in the projected space. This also highlights the inherent difficulty of AUD. Interestingly, when visualizing $\mathbf{F}^{DC}$, as discussed in Appendix D, we observe that the resulting clusters align closely with subject identities, as discussed in the supplementary materials. This supports our hypothesis that $\mathbf{F}^{DC}$ primarily captures temporally invariant, subject-related redundancy.

Facial AU Co-activation and Independence. Figure 7 shows the Pearson correlation heatmap among facial AUs. Red indicates positive correlation, blue indicates negative correlation, and darker shades represent stronger relationships. Detailed dataset-specific analyses are provided in Appendix C. (1) Some AUs show strong positive correlations. AU1 and AU2 have a high correlation ($r = 0.73$), reflecting frequent co-activation during expressions like surprise or focused attention, as both involve brow raising. AU6 and AU12 show moderate positive correlation ($r = 0.20$), consistent with their co-occurrence in genuine smiles. AU15 and AU17 ($r = 0.18$) also show mild correlation, possibly due to their joint involvement in lower-face movements. (2) Some AUs exhibit strong negative correlations. AU4 (Brow Lowerer) is negatively correlated with AU1 and AU2 ($r = -0.31, -0.32$), indicating an antagonistic relationship

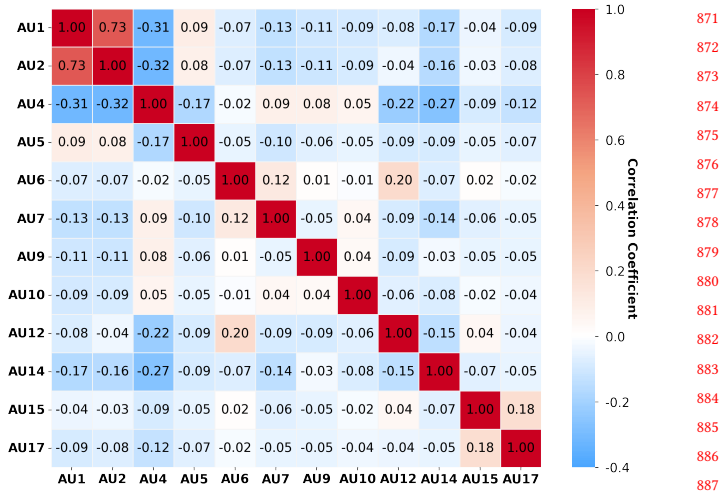


Figure 7: Clustered heatmap of Pearson correlation coefficients among AUs. Color intensity indicates the strength of correlation, with red representing positive correlations and blue representing negative correlations.

between brow lowering and raising. AU4 and AU14 ($r = -0.27$) also rarely co-occur, likely due to their associations with different emotional states (e.g., negative vs. positive/sarcastic). (3) Most AU pairs show weak or no correlation. AUs with $|r| < 0.1$ are considered uncorrelated. For example, AU10 shows little correlation with others, as do AU5, AU9, AU15, and AU17, suggesting high independence in MEs. A correlation network is included in Appendix C. In summary, strong correlations often reflect shared regions or emotional functions, while negative correlations indicate opposing movements. Most AUs are independent, highlighting the flexibility of facial expressions. Key AU relationships like AU1-AU2, AU6-AU12, and AU4-AU1/2 warrant special attention in AUD.

5 Conclusion

In this paper, we present AC4AU, an apex-free framework for Action Unit Detection (AUD) that tackles noise interference, information redundancy, and sample imbalance. A Frequency-aware Redundancy Decomposer (FRD) removes DC components to retain dynamic, emotion-related AC features. An AU-specific Expert Router (AUsER), built on the Mixture-of-Experts paradigm, adaptively routes features to specialized AU predictors for enhanced generalization. Additionally, a face recognition backbone provides robust frame-level encoding without requiring alignment.

Extensive and rigorous experiments under strict LODO protocols demonstrate that AC4AU not only achieves competitive results compared to apex-dependent methods but also ensures fair and reliable evaluations. Besides, we also provide statistical insights into AU dependencies, offering new perspectives on the relationship between local AU activations and global emotional processes. In our future work, we will expand our analysis from AUD to MER tasks within the CD6ME benchmark, incorporating large language models and dynamic network planning to promote fair and reliable comparisons in the micro-expression research community.

References

[1] Xianye Ben, Chen Gong, Tianhuan Huang, Chuanye Li, Rui Yan, and Yujun Li. 2023. Tackling Micro-Expression Data Shortage via Dataset Alignment and Active Learning. *IEEE Trans. Multim.* 25 (2023), 5429–5443. doi:10.1109/TMM.2022.3192727

[2] Xianye Ben, Yi Ren, Junping Zhang, Su-Jing Wang, Kidiyo Kpalma, Weixiao Meng, and Yong-Jin Liu. 2022. Video-based facial micro-expression analysis: A survey of datasets, features and algorithms. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 9 (2022), 5826–5846. doi:10.1109/TPAMI.2021.3067464

[3] Charles S Carver and Michael F Scheier. 1982. Control theory: A useful conceptual framework for personality–social, clinical, and health psychology. *Psychological bulletin* 92, 1 (1982), 111.

[4] Sheng Chen, Yang Liu, Xiang Gao, and Zhen Han. 2018. MobileFaceNets: Efficient CNNs for Accurate Real-Time Face Verification on Mobile Devices. In *Proceedings of the 13th Chinese Conference on Biometric Recognition (Lecture Notes in Computer Science, Vol. 10996)*. Springer, 428–438. doi:10.1007/978-3-319-97909-0_46

[5] Arun Das, Jeffrey Mock, Yufei Huang, Edward J. Golob, and Peyman Najafirad. 2021. Interpretable self-supervised facial micro-expression learning to predict cognitive state and neurological disorders. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence*. AAAI Press, 818–826. doi:10.1609/AAAI.V35I1.16164

[6] Adrian K. Davison, Cliff Lansley, Nicholas Costen, Kevin Tan, and Moi Hoon Yap. 2018. SAMM: A Spontaneous Micro-Facial Movement Dataset. *IEEE Trans. Affect. Comput.* 9, 1 (2018), 116–129. doi:10.1109/TAFFC.2016.2573832

[7] Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos Zafeiriou. 2022. ArcFace: Additive angular margin loss for deep face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 10 (2022), 5962–5979. doi:10.1109/TPAMI.2021.3087709

[8] Zizhao Dong, Gang Wang, Shaoyuan Lu, Jingting Li, Wenjing Yan, and Su-Jing Wang. 2022. Spontaneous facial expressions and micro-expressions coding: From brain to face. *Frontiers in Psychology* 12 (2022), 784834. doi:10.3389/fpsyg.2021.784834

[9] Paul Ekman and W. V. Friesen. 1978. Facial action coding system (FACS): A technique for the measurement of facial actions. *Rivista Di Psichiatria* 47, 2 (1978), 126–138.

[10] Xinqi Fan, Xueli Chen, Mingjie Jiang, Ali Raza Shahid, and Hong Yan. 2023. SelfME: Self-supervised motion learning for micro-expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 13834–13843. doi:10.1109/CVPR52729.2023.01329

[11] Xiqiao Fang, Qingfeng Wu, and Lu Cao. 2024. SPCL-MER: Supervised Prototypical Contrastive Learning for Micro-Expression Recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*. IEEE, 5690–5694. doi:10.1109/ICASSP48485.2024.10448102

[12] Yee Siang Gan, Sze-Teng Liong, Wei-Chuen Yau, Yen-Chang Huang, and Tan Lit Ken. 2019. OFF-ApexNet on micro-expression recognition system. *Signal Process. Image Commun.* 74 (2019), 129–139. doi:10.1016/J.IMAGE.2019.02.005

[13] Yuhong He, Wenchao Liu, Guangyu Wang, Lin Ma, and Haifeng Li. 2024. Enhancing micro-expression analysis performance by effectively addressing data imbalance. In *Proceedings of the 32nd ACM International Conference on Multimedia*. ACM, 11503–11507. doi:10.1145/3664647.3689144

[14] E Tory Higgins. 1997. Beyond pleasure and pain. *American psychologist* 52, 12 (1997), 1280.

[15] Ankith Jain Rakesh Kumar and Bir Bhanu. 2021. Micro-expression classification based on landmark relations With graph attention convolutional network. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*. Computer Vision Foundation / IEEE, 1511–1520. doi:10.1109/CVPRW53098.2021.00167

[16] Ankith Jain Rakesh Kumar and Bir Bhanu. 2022. Three Stream Graph Attention Network using Dynamic Patch Selection for the classification of micro-expressions. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 2475–2484. doi:10.1109/CVPRW56347.2022.00277

[17] Ankith Jain Rakesh Kumar and Bir Bhanu. 2023. Relational edge-node Graph attention network for classification of micro-expressions. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 5819–5828. doi:10.1109/CVPRW59228.2023.00618

[18] Hanting Li, Mingzhe Sui, Zhaoqing Zhu, and Feng Zhao. 2022. MMNet: Muscle motion-guided network for micro-expression recognition. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*. ijcai.org, 1074–1080. doi:10.24963/IJCAI.2022/150

[19] Jingting Li, Zizhao Dong, Shaoyuan Lu, Su-Jing Wang, Wen-Jing Yan, Yinhuan Ma, Ye Liu, Changbing Huang, and Xiaolan Fu. 2023. CAS(ME)^3: A Third Generation Facial Spontaneous Micro-Expression Database With Depth Information and High Ecological Validity. *IEEE Trans. Pattern Anal. Mach. Intell.* 45, 3 (2023), 2782–2800. doi:10.1109/TPAMI.2022.3174895

[20] Jingting Li, Su-Jing Wang, Yong Wang, Haoliang Zhou, and Xiaolan Fu. 2025. Parallel Spatiotemporal Network to recognize micro-expression. *Neurocomputing* 636 (2025), 129891. doi:10.1016/J.NEUCOM.2025.129891

[21] Jingting Li, Haoliang Zhou, Yu Qian, Zizhao Dong, and Su-Jing Wang. 2025. Micro-expression recognition using dual-view self-supervised contrastive learning with intensity perception. *Neurocomputing* 619 (2025), 129142. doi:10.1016/J.NEUCOM.2024.129142

[22] Xiaobai Li, Shiyang Cheng, Yante Li, Muzammil Behzad, Jie Shen, Stefanos Zafeiriou, Maja Pantic, and Guoying Zhao. 2023. 4DME: A spontaneous 4D micro-expression dataset With multimodalities. *IEEE Trans. Affect. Comput.* 14, 4 (2023), 3031–3047. doi:10.1109/TAFFC.2022.3182342

[23] Xiaobai Li, Xiaopeng Hong, Antti Moilanen, Xiaohua Huang, Tomas Pfister, Guoying Zhao, and Matti Pietikäinen. 2018. Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods. *IEEE Trans. Affect. Comput.* 9, 4 (2018), 563–577.

[24] Yante Li, Xiaohua Huang, and Guoying Zhao. 2021. Micro-expression action unit detection with spatial and channel attention. *Neurocomputing* 436 (2021), 221–231. doi:10.1016/J.NEUCOM.2021.01.032

[25] Yante Li, Jinsheng Wei, Yang Liu, Janne Kauttonen, and Guoying Zhao. 2022. Deep Learning for Micro-Expression Recognition: A Survey. *IEEE Trans. Affect. Comput.* 13, 4 (2022), 2028–2046. doi:10.1109/TAFFC.2022.3205170

[26] Sze-Teng Liong, Yee Siang Gan, John See, Huai-Qian Khor, and Yen-Chang Huang. 2019. Shallow triple stream three-dimensional CNN (STSTNet) for micro-expression recognition. In *Proceedings of the 14th IEEE International Conference on Automatic Face & Gesture Recognition*. IEEE, 1–5. doi:10.1109/FG.2019.8756567

[27] Sze-Teng Liong, John See, KokSheik Wong, and Raphael C.-W. Phan. 2018. Less is more: Micro-expression recognition from video using apex frame. *Signal Process. Image Commun.* 62 (2018), 82–92. doi:10.1016/J.IMAGE.2017.11.006

[28] Yong-Jin Liu, Jinkai Zhang, Wen-Jing Yan, Sujing Wang, Guoying Zhao, and Xiaolan Fu. 2016. A main directional mean optical flow feature for spontaneous micro-expression recognition. *IEEE Trans. Affect. Comput.* 7, 4 (2016), 299–310. doi:10.1109/TAFFC.2015.2485205

[29] Xuan-Bac Nguyen, Chi Nhan Duong, Xin Li, Susan Gauch, Han-Seok Seo, and Khoa Luu. 2023. Micron-BERT: BERT-based facial micro-expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 1482–1492. doi:10.1109/CVPR52729.2023.00149

[30] Wenfeng Qin, Bochao Zou, Xin Li, Weiping Wang, and Huimin Ma. 2023. Micro-Expression Spotting with Face Alignment and Optical Flow. In *Proceedings of the 31st ACM International Conference on Multimedia*. ACM, 9501–9505. doi:10.1145/3581783.3612853

[31] Shyam Sundar Rajagopalan, O. V. Ramana Murthy, Roland Goecke, and Agata Rozga. 2015. Play with me - Measuring a child's engagement in a social interaction. In *11th IEEE International Conference and Workshops on Automatic Face & Gesture Recognition*. IEEE Computer Society, 1–8. doi:10.1109/FG.2015.7163129

[32] Nabeel Seedat, Fergus Imrie, and Mihaela van der Schaar. 2024. Navigating Data-Centric Artificial Intelligence With DC-Check: Advances, Challenges, and Opportunities. *IEEE Trans. Artif. Intell.* 5, 6 (2024), 2589–2603. doi:10.1109/TAI.2023.3345805

[33] Pei-Sze Tan, Sailaja Rajanala, Arghya Pal, Shu-Min Leong, Raphaël C.-W. Phan, and Huey Fang Ong. 2024. Causally Uncovering Bias in Video Micro-Expression Recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 5790–5794. doi:10.1109/ICASSP48485.2024.10447476

[34] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9, 11 (2008), 2579–2605.

[35] Tuomas Varanka, Yante Li, Wei Peng, and Guoying Zhao. 2024. Data leakage and evaluation issues in micro-expression analysis. *IEEE Trans. Affect. Comput.* 15, 1 (2024), 186–197. doi:10.1109/TAFFC.2023.3265063

[36] Tuomas Varanka, Wei Peng, and Guoying Zhao. 2021. Micro-Expression Recognition with Noisy Labels. In *Human Vision and Electronic Imaging 2021*. Society for Imaging Science and Technology, 157–1–157–8. doi:10.2352/ISSN.2470-1173.2021.11.HVEI-157

[37] Tuomas Varanka, Wei Peng, and Guoying Zhao. 2023. Learnable eulerian dynamics for micro-expression action unit detection. In *Proceedings of the 22nd Scandinavian Conference on Image Analysis (Lecture Notes in Computer Science, Vol. 13886)*. Springer, 385–400. doi:10.1007/978-3-031-31438-4_26

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Curran Associates, Inc., 5998–6008.

[39] Feifan Wang, Yuan Zong, Jie Zhu, Mengting Wei, Xiaolin Xu, Cheng Lu, and Wenming Zheng. 2024. Progressively Learning from Macro-Expressions for Micro-Expression Recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 4390–4394. doi:10.1109/ICASSP48485.2024.10446028

[40] Mengting Wei, Xingxun Jiang, Wenming Zheng, Yuan Zong, Cheng Lu, and Jiateng Liu. 2023. CMNet: Contrastive magnification network for micro-expression recognition. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence*. AAAI Press, 119–127. doi:10.1609/AAAI.V37I1.25083

[41] Hong-Xia Xie, Ling Lo, Hong-Han Shuai, and Wen-Huang Cheng. 2023. An Overview of Facial Micro-Expression Analysis: Data, Methodology and Challenge. *IEEE Trans. Affect. Comput.* 14, 3 (2023), 1857–1875. doi:10.1109/TAFFC.2022.3143100

[42] Shiting Xu, Zhiheng Zhou, and Junyuan Shang. 2022. Asymmetric adversarial-based feature disentanglement learning for cross-database micro-expression recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*. ACM, 5342–5350. doi:10.1145/3503161.3548435

[43] Wen-Jing Yan, Qi Wu, Yong-Jin Liu, Sujing Wang, and Xiaolan Fu. 2013. CASME database: A dataset of spontaneous micro-expressions collected from neutralized faces. In *Proceedings of the 10th IEEE International Conference on Automatic Face & Gesture Recognition*. IEEE Computer Society, 1–7. doi:10.1109/FG.2013.6553799

[44] Wen-Jing Yan, Xiaobai Li, Su-Jing Wang, Guoying Zhao, Yong-Jin Liu, Yu-Hsin Chen, and Xiaolan Fu. 2014. CASME II: An improved spontaneous micro-expression database and the baseline evaluation. *PloS one* 9, 1 (2014), e86041. doi:10.1371/journal.pone.0086041

[45] Wen-Jing Yan, Qi Wu, Jing Liang, Yu-Hsin Chen, and Xiaolan Fu. 2013. How fast are the leaked facial expressions: The duration of micro-expressions. *Journal of nonverbal behavior* 37 (2013), 217–230. doi:10.1007/s10919-013-0159-8

[46] Jun Yu, Yaohui Zhang, Gongpeng Zhao, Peng He, Zerui Zhang, Zhongpeng Cai, Qingsong Liu, Jianqing Sun, and Jiaen Liang. 2024. Micro-Expression Spotting Based on Optical Flow Feature with Boundary Calibration. In *Proceedings of the 32nd ACM International Conference on Multimedia*. ACM, 11490–11496. doi:10.1145/3664647.3689142

[47] Jun Yu, Gongpeng Zhao, Yaohui Zhang, Peng He, Zerui Zhang, Zhao Yang, Qingsong Liu, Jianqing Sun, and Jiaen Liang. 2024. Temporal-informative adapters in VideoMAE V2 and multi-scale feature fusion for micro-expression spotting-then-recognize. In *Proceedings of the 32nd ACM International Conference on Multimedia*. ACM, 11484–11489. doi:10.1145/3664647.3689141

[48] Zhijun Zhai, Jianhui Zhao, Chengjiang Long, Wenju Xu, Shuangjiang He, and Huijuan Zhao. 2023. Feature representation learning with adaptive displacement generation and transformer fusion for micro-expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 22086–22095. doi:10.1109/CVPR52729.2023.02115

[49] Fuli Zhang, Yu Liu, Xiaoling Yu, Zhichen Wang, Qi Zhang, Jing Wang, and Qionghua Zhang. 2025. Towards facial micro-expression detection and classification using modified multimodal ensemble learning approach. *Inf. Fusion* 115 (2025), 102735. doi:10.1016/J.INFFUS.2024.102735

[50] Jiahao Zhang, Feng Liu, and Aimin Zhou. 2021. Off-TANet: A lightweight neural micro-expression recognizer with optical flow features and integrated attention mechanism. In *Proceedings of the 18th Pacific Rim International Conference on Artificial Intelligence (Lecture Notes in Computer Science, Vol. 13031)*. Springer, 266–279. doi:10.1007/978-3-030-89188-6_20

[51] Lijun Zhang, Yifan Zhang, Xinzhi Sun, Weicheng Tang, Xiaomeng Wang, and Zhanshan Li. 2025. Micro-expression recognition based on direct learning of graph structure. *Neurocomputing* 619 (2025), 129135. doi:10.1016/J.NEUCOM.2024.129135

[52] Xing Zhang, Lijun Yin, Jeffrey F. Cohn, Shaun J. Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M. Girard. 2014. BP4D-Spontaneous: A high-resolution spontaneous 3D dynamic facial expression database. *Image Vis. Comput.* 32, 10 (2014), 692–706. doi:10.1016/J.IMAVIS.2014.06.002

[53] Guoying Zhao and Xiaobai Li. 2019. Automatic micro-expression analysis: Open challenges. *Frontiers in psychology* 10 (2019), 1833. doi:10.3389/fpsyg.2019.01833

[54] Guoying Zhao and Matti Pietikäinen. 2007. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans. Pattern Anal. Mach. Intell.* 29, 6 (2007), 915–928. doi:10.1109/TPAMI.2007.1110