Preprocessing Paradox in Vision Transformer-Based Micro-Expression Recognition

1st Muhammad Taufiq Al Fikri
School of Computing
Telkom University
Bandung, Indonesia
taufiqafk@student.telkomuniversity.ac.id

2nd Kurniawan Nur Ramadhani
Center of Excellence of Artificial Intelligence and Learning Optimization,
School of Computing, Telkom University
Bandung, Indonesia
kurniawannr@telkomuniversity.ac.id

***Abstract*—Micro-expression recognition faces significant challenges from subtle facial movements lasting 40-500 milliseconds and severely limited training data. This study systematically compares three Vision Transformer architectures—ViT, Swin Transformer, and PoolFormer—representing distinct token mixing strategies across dual preprocessing methodologies on the CASME II dataset. Contrary to conventional computer vision assumptions, we discover a counterintuitive preprocessing paradox: systematic face-aware preprocessing with superior geometric quality (100% face centering) degraded temporal performance by 18.4-20.5% compared to minimally processed raw images. On small datasets (201 training samples), clean preprocessing paradoxically enables models to overfit to superficial patterns and preprocessing-specific artifacts, while "noisy" raw images with spatial variability enforce robust invariance learning through natural augmentation. Additionally, attention-free PoolFormer demonstrated unique temporal robustness unsuited for single frames but excelling at sequential modeling, achieving best overall performance (macro F1=0.4762) with remarkable 67.4% improvement as temporal density increased from single apex frames to dense multi-frame sampling, contrasting sharply with ViT's 40.5% degradation despite superior peak-frame performance. These findings fundamentally challenge preprocessing optimization assumptions for resource-constrained scenarios and establish PoolFormer's attention-free architecture as particularly promising for sequential micro-expression modeling. We provide the first rigorous 7-category CASME II benchmark with comprehensive Vision Transformer evaluation under extreme 49.5:1 class imbalance, offering critical insights for advancing micro-expression recognition research in small-data regimes.**

***Index Terms*—micro-expression recognition, Vision Transformer, preprocessing paradox, PoolFormer, temporal aggregation, CASME II**

I. Introduction

Micro-expressions are involuntary facial expressions lasting 40-500 milliseconds that reveal genuine emotions during concealment attempts, serving as critical indicators in psychological analysis, security screening, and deception detection [1], [2]. Their fleeting nature and subtle intensity create significant automated recognition challenges, while manual analysis remains time-consuming and requires extensively trained experts [3].

Convolutional Neural Networks' localized operations with limited receptive fields constrain long-range dependency modeling essential for subtle micro-expression discrimination [4], [5]. Vision Transformers address this through self-attention mechanisms capturing global relationships, with recent architectures demonstrating promising results [5]–[8]. However, comprehensive evaluations comparing distinct token mixing strategies—global attention, hierarchical windows, and attention-free pooling—remain unexplored. Furthermore, systematic preprocessing impact on small-scale datasets lacks rigorous validation despite conventional optimization assumptions.

This study systematically investigates three Vision Transformer architectures—ViT (global self-attention), Swin Transformer (hierarchical windowed attention), and PoolFormer (attention-free pooling) [9]—across three temporal sampling strategies on CASME II [1]. Our investigation reveals a counterintuitive preprocessing paradox where face-aware optimization degraded performance by 20.5% compared to raw images on small datasets, while PoolFormer demonstrated unique temporal robustness achieving best overall performance through consistent improvement with temporal density, contrasting sharply with ViT's severe degradation.

Our contributions include: (1) discovering the preprocessing paradox challenging conventional assumptions, (2) providing first comprehensive Vision Transformer comparison on 7-category CASME II with 49.5:1 class imbalance, (3) identifying PoolFormer's temporal robustness establishing future sequential modeling directions, and (4) developing Multi-Frame Sampling strategy achieving 10.25× density expansion. These findings establish that architectural simplicity with temporal robustness may outperform complex attention mechanisms in resource-constrained scenarios, while the preprocessing paradox highlights critical dataset-scale considerations for methodology design.

II. Materials and Methods

A. Dataset and Experimental Design

This research utilized CASME II [1], comprising 255 video sequences from 26 Chinese participants captured at 200fps with 640×480 resolution. The dataset exhibits severe class imbalance with a 49.5:1 ratio between Others (99 samples) and Fear (2 samples), with intermediate classes including Disgust (63), Happiness (32), Repression (27), Surprise (25), and Sadness (7). Stratified splitting maintained class distribution across training (78.8%, 201 videos), validation (10.2%, 26 videos), and test (11.0%, 28 videos) sets.

We conducted experiments under dual methodologies to evaluate preprocessing impact (Table I). Methodology 1 (M1) employed minimal preprocessing with center-cropped raw frames at 384×384 RGB resolution, preserving original data characteristics. Methodology 2 (M2) implemented systematic face-aware preprocessing through a five-stage pipeline: Dlib face detection [10], bounding box expansion (+20 pixels), direct cropping, resizing to 224×224, and grayscale conversion. M2 achieved superior geometric quality with 100% face centering compared to M1's 84%, though information density reduced by 88.7%.

TABLE I
Dataset and Methodology Comparison

Aspect	M1 (Raw)	M2 (Face)
Resolution	384×384	224×224
Color	RGB	Grayscale
Face Center	84%	100%
Info. Density	442K	50K
Training Samples by Phase:		
AF (1×)	201	201
KFS (3×)	603	603
MFS (10.25×)	2,061	2,061

Each methodology was evaluated across three temporal sampling phases illustrated in Figure 1. Apex Frame (AF) used single peak-intensity frames yielding 201 training samples. Key Frame Sequence (KFS) captured onset-apex-offset transitions with 603 frames (3× expansion) and video-level predictions through majority voting. Multi-Frame Sampling (MFS) exploited CASME II metadata through adaptive windowing around key frames (onset: +[0,1,2,3], apex: [-2,-1,0,1,2], offset: [-3,-2,-1,0]), averaging 13 frames per video with 10.25× expansion to 2,061 samples.

B. Model Architectures

Three Vision Transformer architectures representing distinct token mixing strategies were selected based on preliminary screening (Table II). ViT-patch32, based on the original Vision Transformer architecture [11], employs full global self-attention where every 32×32 patch attends to all others simultaneously, yielding 49 tokens at 224×224 resolution with O(N²) complexity. While Micron-BERT [5] adapted this for micro-expressions, our implementation uses standard ViT to isolate architectural effects.

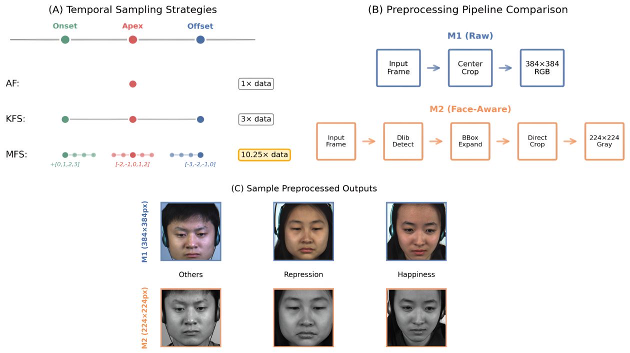

Fig. 1. Experimental framework with temporal sampling strategies and preprocessing pipelines.

TABLE II
Architecture Characteristics Comparison

Model	Token Mixing	Strength
ViT-patch32	Global attention	Peak intensity
SwinT-base	Hierarchical window	Multi-scale
PoolFormer	Attention-free pool	Temporal robust

SwinT-base [12] implements hierarchical shifted window attention, balancing local detail capture within windows and global context through cross-window connections via shifting mechanisms. This O(N) complexity architecture provides multi-scale feature extraction naturally suited for facial expression analysis at different spatial granularities, as demonstrated in hybrid approaches [6].

PoolFormer-m36/m48 [9] utilizes attention-free pooling operations, replacing learned attention with fixed pooling kernels for deterministic spatial aggregation. With 56M (m36) and 73M (m48) parameters respectively, these models achieve O(N) complexity through architectural simplicity, potentially conferring robustness advantages on small datasets by avoiding attention overfitting. All models initialized with ImageNet weights for transfer learning benefits.

C. Training and Evaluation

To address class imbalance, we compared class-weighted CrossEntropy (with inverse frequency weighting) against Focal Loss [13] (using alpha-weighting and $\gamma = 2.0$). Loss selection was architecture-specific, with ViT and PoolFormer-m36 favoring Focal Loss while SwinT and PoolFormer-m48 preferred CrossEntropy based on validation performance. Training employed the AdamW optimizer with a learning rate of 5×10^{-5} and weight decay of 0.001 for 50 epochs, utilizing early stopping (patience=3-5) based on validation macro F1-score. Hyperparameter optimization via Optuna tuned drop path rate (0.05-0.3), dropout (0.0-0.2), and layer configurations. Batch sizes varied dynamically ($16 - 32$ for M1, $4 - 16$ for M2) due to resolution and phase constraints. Experiments were conducted on Google Colab's NVIDIA L4 GPU (22GB VRAM) using PyTorch.

Performance assessment relied on macro F1-score as the primary evaluation criterion, treating all seven emotion classes equally despite severe imbalance to ensure a fair and realistic model capability evaluation (unlike common class-merging approaches). To establish statistical reliability on the small test set ($n = 28$), 95% bootstrap confidence intervals (CI) were calculated using 1,000 resampling iterations with replacement and the percentile method. The macro F1 calculation only included emotion classes present in the test set, excluding the Fear class (zero test samples), thus providing meaningful statistical interpretation under class imbalance conditions.

III. Results and Discussion

A. Preprocessing Paradox: Performance Degradation with Quality

Our investigation reveals a counterintuitive finding challenging conventional preprocessing assumptions. Despite achieving superior geometric localization (100% face centering versus 84% for raw images), systematic face-aware preprocessing (M2) consistently degraded performance compared to minimally processed raw images (M1) across temporal scenarios (Table III).

TABLE III
Main Results: Best Models per Phase with Statistical Validation

Phase	Model	M1 F1 (95% CI)	M2 F1
AF (1×)	ViT-p32	0.4235 [0.22, 0.68]	0.4229
KFS (3×)	Pool-m48	0.3974	0.3241
MFS (10.25×)	Pool-m36	**0.4762 [0.26, 0.71]**	0.3785

Note: Bootstrap confidence intervals (1,000 iterations) provided for best-performing single-frame (AF) and best overall (MFS) models. KFS performance excluded from statistical validation due to intermediate status.

Degradation severity intensified with temporal expansion. While single-frame Apex Frame (AF) evaluation showed minimal difference (M1: 0.4235 versus M2: 0.4229 macro F1, -0.1%), temporal aggregation revealed striking divergence: Key Frame Sequence (KFS) exhibited 18.4% degradation (M1: 0.3974 versus M2: 0.3241), escalating to 20.5% in Multi-Frame Sampling (MFS) phase (M1: 0.4762 versus M2: 0.3785). Figure 2 demonstrates how M2's superior face centering resulted in worse performance despite better geometric localization, with preprocessing reducing information density by 88.7% through resolution reduction and grayscale conversion.

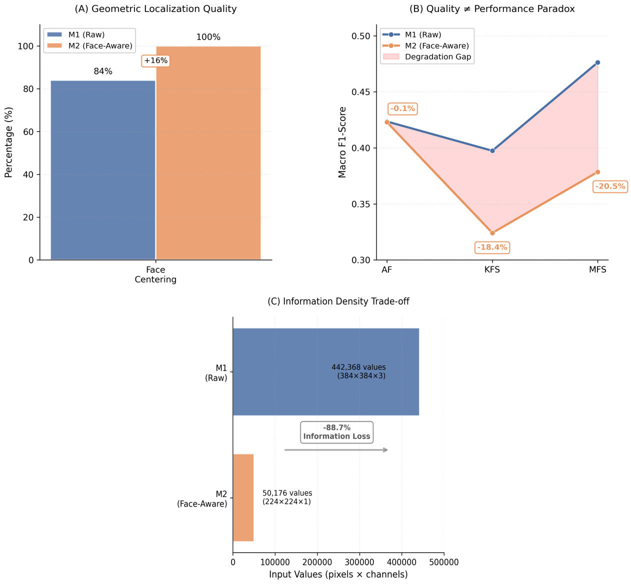

Fig. 2. Preprocessing paradox: quality improvement versus performance degradation.

Bootstrap analysis revealed substantial uncertainty (ViT AF 95% CI: [0.22, 0.68], PoolFormer MFS 95% CI: [0.26, 0.71]), with confidence interval widths of 109% and 94.5% reflecting small test set limitations. Nevertheless, consistent M1 superiority across all temporal phases—with M1 point estimates exceeding M2 by margins substantially larger than confidence interval overlaps—establishes the preprocessing paradox as a robust finding rather than sampling artifact.

The M2 pipeline's resolution decrease (384×384→224×224, -65.3%) and grayscale conversion eliminated potentially discriminative color cues from blood flow changes, with entropy-based attention research [7] and composite database findings [14] demonstrating multi-modal features' importance for subtle expression recognition.

We hypothesize this paradox stems from overfitting dynamics specific to severe data scarcity (201 training samples). Clean, geometrically-precise M2 images enable models to memorize preprocessing-specific artifacts rather than learning robust expression features, while ”noisy” M1 images containing spatial variability enforce invariance learning. Supporting evidence includes degradation correlating with training set size expansion, simpler architectures showing less sensitivity, and temporal scenarios requiring cross-frame generalization suffering most severely, consistent with observations in other small-scale vision tasks [15].

B. Architecture-Specific Temporal Behaviors

Vision Transformer variants demonstrated fundamentally different responses to temporal aggregation, revealing architecture-dependent suitability for sequential micro-expression modeling (Table IV). ViT-patch32 achieved the strongest single-frame performance (M1 AF: 0.4235 macro F1) through global self-attention [11], but experienced severe temporal degradation (KFS: 0.2520, -40.5%) when aggregating conflicting signals. In contrast, Swin Transformer maintained consistent, moderate performance (M1: 0.3820-0.4075 range) across phases using hierarchical shifted window attention [12], balancing local and global features but lacking specialized advantages.

TABLE IV
TEMPORAL EVOLUTION PATTERNS

Model	**AF**	**KFS**	**MFS**
ViT-p32	0.4235	0.2520↓	0.3285↑
SwinT	0.4075	0.3820↓	0.3282↓
Pool-m36	0.2844	0.3505↑	**0.4762↑**

The performance trajectory of all models is visualized in Figure 3. The attention-free PoolFormer demonstrated a unique temporal aggregation emergence, showing a consistent upward trajectory: AF (0.2844) → MFS (0.4762, $+67.4\%$ total). This robustness stems from the pooling mechanism's fixed kernels providing deterministic spatial integration, making it less sensitive to temporal noise than learned attention weights, a finding aligned with MetaFormer research [9]. The smaller m36 variant (56M parameters) also outperformed m48 (73M) in dense sampling (0.4762 versus 0.4038), indicating architectural simplicity benefits small datasets.

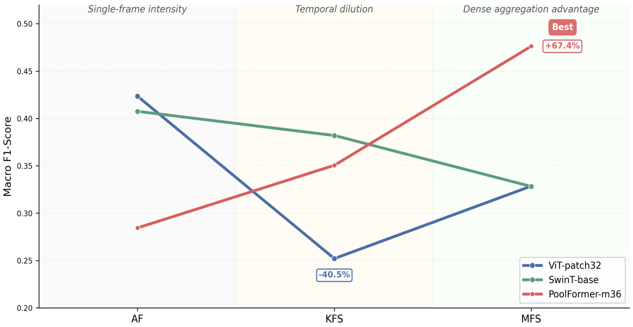

Fig. 3. Architecture-specific temporal aggregation patterns across three phases.

C. Per-Class Performance Analysis

Emotion-wise analysis using best-performing PoolFormer-m36 revealed distinct difficulty hierarchies and preprocessing sensitivities (Table V). Figure 4 shows the confusion matrix highlighting primary error patterns between emotion classes.

TABLE V
PER-CLASS F1-SCORES FOR BEST PERFORMING MODELS

Emotion	**ViT M1**	**ViT M2**	**Pool M1**	**Pool M2**	**Samples**
Others	0.636	0.545↓	0.571	0.455↓	10
Disgust	0.500	0.714↑	0.533	0.667↑	7
Happiness	0.333	0.333	0.286	0.250↓	4
Repression	0.571	0.444↓	0.667	0.400↓	3
Surprise	0.500	0.500	0.800	0.500↓	3
Sadness	0.000	0.000	0.000	0.000	1
Fear	—	—	—	—	0

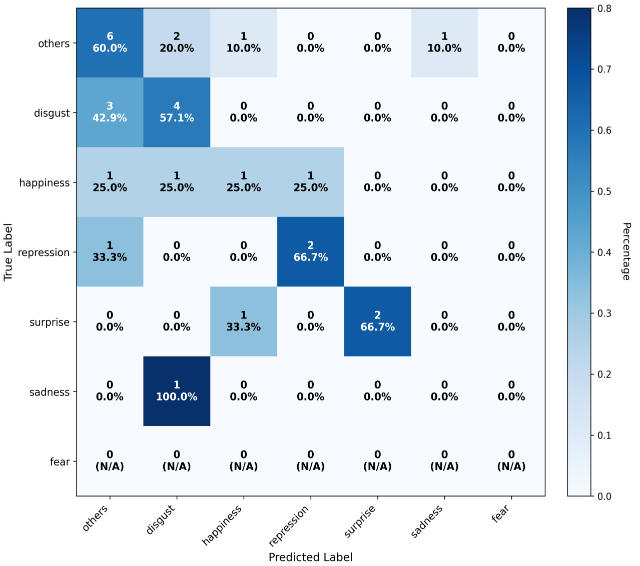

Fig. 4. Confusion matrix for PoolFormer-m36 with M1 preprocessing.

Surprise emerged most recognizable (average F1: 0.612) due to its distinctive patterns, while Happiness proved most challenging among adequately-sampled classes (F1: 0.282), potentially due to variability between genuine and polite smiles. Mid-tier emotions—Repression (0.580), Disgust (0.567), and Others (0.520)—showed moderate recognition. Fear and Sadness demonstrated fundamental limitations (F1: 0.000) due to extreme scarcity, suggesting that combining multiple datasets might be necessary to address this imbalance [8], [14]. Confusion was primarily observed between Others-Disgust pairs (5 bidirectional errors), reflecting perceptual similarity in neutral-negative boundaries and shared action units according

TABLE VI
CONTEXTUAL COMPARISON WITH STATE-OF-THE-ART METHODS

Method	**Year**	**Approach**	**Protocol**	**Metric**	**Score**	**Notes**
AUMEs [16]	2025	Dual 3D-CNN + AU	5-category	UF1	0.8880	Multi-task, Focal Loss
MobileViT [17]	2022	Transfer + SVM	LOSO	UF1	0.7251	Lightweight
Faciaformer [6]	2023	CNN+LSTM+ViT	Mixed	Acc	0.9112	Score fusion ensemble
Ours (Pool-m36)	2025	Attention-free ViT	**7-category**	**F1**	**0.4762**	Full imbalance challenge
Ours (ViT-p32)	2025	Global attention	**7-category**	**F1**	**0.4235**	Apex frame baseline

to FACS theory [18]. Secondary confusion between Repression-Others (3 errors) highlighted the challenge of distinguishing suppressed expressions from neutral states [16].

Regarding Emotion-Specific Preprocessing Effects, Disgust paradoxically benefited from M2 preprocessing (ViT: +42.9%, PoolFormer: +25.0%), suggesting facial structure emphasis enhanced disgust-specific action unit detection. Conversely, Repression suffered severe M2 degradation (ViT: -22.2%, PoolFormer: -40.0%), potentially due to the loss of subtle color cues critical for detecting suppressed expressions where conscious inhibition minimizes overt movements. This supports findings that preserving multi-modal information improves subtle expression detection [19].

D. Computational Efficiency

Inference speed analysis demonstrated practical deployment viability across all architectures. ViT-patch32 achieved fastest processing (19.67ms/sample, 51 FPS) through larger patch size reducing token count. PoolFormer exhibited resolution-dependent latency: M2's 224×224 input enabled 28.26ms processing versus M1's 384×384 requiring 74.56ms. All models exceeded real-time thresholds (>13 FPS), supporting practical applications in security screening and clinical interviews as explored in recent acoustic-visual environment studies [20].

E. Comparison with State-of-the-Art

Direct comparison with existing literature requires acknowledging fundamental protocol differences (Table VI). Recent approaches employ various simplification strategies: AUMEs [16] uses 5-category classification with AU-based features reaching UF1=0.888, Lightweight ViT [17] achieved UF1=0.725 through transfer learning, while composite database approaches [14] merge multiple datasets for improved generalization. Our 7-category evaluation deliberately confronts complete dataset complexity including 49.5:1 imbalance. While absolute macro F1 (0.4762) appears lower, we address a fundamentally harder problem requiring generalization across severely underrepresented classes.

F. Critical Insights and Implications

Our findings reveal four key implications. First, optimal preprocessing depends on dataset size: <500 samples favor minimal preprocessing, 500-1,000 require validation, >1,000 benefit from systematic preprocessing. Second, architecture selection should prioritize temporal robustness (PoolFormer: +67.4%) over peak accuracy (ViT: -40.5%). Third, loss function selection must couple with architecture—ViT benefits from Focal Loss [13] while PoolFormer shows phase-dependent preferences. Fourth, PoolFormer establishes attention-free architectures as promising for temporal modeling, with potential LSTM/GRU enhancement [6]. These insights challenge conventional wisdom favoring sophisticated preprocessing and attention mechanisms, establishing architectural simplicity and information preservation as critical factors in resource-constrained scenarios.

IV. CONCLUSION

This systematic investigation of Vision Transformer architectures for micro-expression recognition on the challenging 7-category CASME II benchmark [1] yielded critical insights by challenging conventional assumptions and establishing promising research directions. Our comprehensive evaluation delivered four key contributions: (1) the discovery of a preprocessing paradox, where systematic face-aware optimization paradoxically degraded temporal performance by 18.4%-20.5% compared to raw images; (2) the identification of the attention-free PoolFormer as uniquely suited for temporal modeling, achieving the best overall performance (macro $F1 = 0.4762$) through consistent improvement with temporal density ($+67.4\%$), in sharp contrast to ViT (-40.5% degradation); (3) the first comprehensive Vision Transformer comparison on the full 7-category CASME II under extreme 49.5:1 class imbalance; and (4) the demonstration of architecture-dependent responses to temporal augmentation via our Multi-Frame Sampling strategy ($10.25\times$ training expansion).

Several limitations contextualize these findings and inform future work. Given the small size of the training dataset (n = 201) and the extreme scarcity of the Fear and Sadness classes (rendering their evaluation statistically unreliable), the preprocessing paradox may be specific to severe data scarcity. This necessitates cross-dataset validation on larger resources like CAS(ME)³ [21] to explore whether depth information mitigates the paradox and to determine precise dataset-size thresholds where optimization transitions from harmful to beneficial (e.g., around 500-1,000 samples). Furthermore, future research should prioritize PoolFormer-based sequential modeling integrating LSTM/GRU modules [6] and explore hybrid architectures combining PoolFormer's temporal aggregation with ViT's peak intensity capture to leverage complementary strengths. Bootstrap confidence intervals for best-performing models quantified this uncertainty, with 95

This work challenges conventional preprocessing wisdom while identifying attention-free architectures as uniquely suited for resource-constrained micro-expression recognition. The preprocessing paradox highlights that what works for large-scale vision tasks may harm performance on specialized small-scale datasets. PoolFormer's emergence establishes clear direction for advancing sequential micro-expression modeling toward practical applications requiring both accuracy and computational efficiency. These insights provide valuable guidance for future micro-expression recognition research, particularly as the field progresses toward larger, more diverse datasets like CAS(ME)³.

References

[1] W.-J. Yan, X. Li, S.-J. Wang, G. Zhao, Y.-J. Liu, Y.-H. Chen, and X. Fu, "Casme ii: An improved spontaneous micro-expression database and the baseline evaluation," *PloS one*, vol. 9, no. 1, p. e86041, 2014.

[2] H. Pan, L. Xie, Z. Wang, B. Liu, M. Yang, and J. Tao, "Review of micro-expression spotting and recognition in video sequences," *Virtual Reality & Intelligent Hardware*, vol. 3, no. 1, pp. 1–17, 2021.

[3] H. Guerdelli, C. Ferrari, W. Barhoumi, H. Ghazouani, and S. Berretti, "Macro-and micro-expressions facial datasets: A survey," *Sensors*, vol. 22, no. 4, p. 1524, 2022.

[4] F. Zhang and L. Chai, "A review of research on micro-expression recognition algorithms based on deep learning," *Neural Computing and Applications*, vol. 36, no. 29, pp. 17787–17828, 2024.

[5] X.-B. Nguyen, C. N. Duong, X. Li, S. Gauch, H.-S. Seo, and K. Luu, "Micron-bert: Bert-based facial micro-expression recognition," in *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pp. 1482–1492, 2023.

[6] Y. Zheng and E. Blasch, "Facial micro-expression recognition enhanced by score fusion and a hybrid model from convolutional lstm and vision transformer," *Sensors*, vol. 23, no. 12, p. 5650, 2023.

[7] Y. Zhang, W. Lin, Y. Zhang, J. Xu, and Y. Xu, "Leveraging vision transformers and entropy-based attention for accurate micro-expression recognition," *Scientific Reports*, vol. 15, no. 1, p. 13711, 2025.

[8] J. Lin, C. Guo, and X. Wu, "Temporal-polar dynamics: Elevating event-based micro-expression recognition," *IEEE Access*, 2024.

[9] W. Yu, M. Luo, P. Zhou, C. Si, Y. Zhou, X. Wang, J. Feng, and S. Yan, "Metaformer is actually what you need for vision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[10] N. A. M. Amin, N. N. A. Sjarif, and S. S. Yuhaniz, "A comparison study of facial feature extraction using mtcnn, retinaface and dlib face detector for personality traits recognition," *Malaysian Journal of Computer Science*, vol. 38, no. 2, 2025.

[11] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[12] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

[13] X. Li, C. Lv, W. Wang, G. Li, L. Yang, and J. Yang, "Generalized focal loss: Towards efficient representation learning for dense object detection," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 3, pp. 3139–3153, 2022.

[14] N. A. Ab Razak and S. Sahran, "Lightweight micro-expression recognition on composite database," *Applied Sciences*, vol. 13, no. 3, p. 1846, 2023.

[15] A.-L. Cîrneanu, D. Popescu, and D. Iordache, "New trends in emotion recognition using image analysis by neural networks, a systematic review," *Sensors*, vol. 23, no. 16, p. 7092, 2023.

[16] H. Shi, Y. Wang, R. Wang, and D. Liu, "Aumes: Au detection-based dual-stream multi-task 3dcnn for micro-expression recognition," *Neural Processing Letters*, vol. 57, no. 1, p. 18, 2025.

[17] Y. Liu, Y. Li, X. Yi, Z. Hu, H. Zhang, and Y. Liu, "Lightweight vit model for micro-expression recognition enhanced by transfer learning," *Frontiers in Neurorobotics*, vol. 16, p. 922761, 2022.

[18] E. Friesen and P. Ekman, "Facial action coding system: a technique for the measurement of facial movement," *Palo Alto*, vol. 3, no. 2, p. 5, 1978.

[19] Z. Xie and H. Chang, "Micro-expression spotting based on multi-modal hierarchical semantic-guided deep fusion model," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, IEEE, 2025.

[20] G. Liu, P. Hu, H. Zhong, Y. Yang, J. Sun, Y. Ji, J. Zou, H. Zhu, and S. Hu, "Effects of the acoustic-visual indoor environment on relieving mental stress based on facial electromyography and micro-expression recognition," *Buildings*, vol. 14, no. 10, p. 3122, 2024.

[21] J. Li, Z. Dong, S. Lu, S.-J. Wang, W.-J. Yan, Y. Ma, Y. Liu, C. Huang, and X. Fu, "Cas (me) 3: A third generation facial spontaneous micro-expression database with depth information and high ecological validity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 2782–2800, 2022.