ISSN 2616-5775 Vol. 9, Issue 6: 37-46, DOI: 10.25236/AJCIS.2026.090606

Micro-Expression Recognition Network Based on Attention Mechanism and Dual-Channel Fusion

Ziyan Fu[1,a], Jie Ying[1,b,*], Jingyun Yang[1,c], Le Fu[2,d]

[1]School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai, China
[2]First Maternity and Infant Hospital, Tongji University, Shanghai, China
[a]3285807082@qq.com, [b,]yingjsh@163.com, [c]2413530329@st.usst.edu.cn, [d]fule0125@qq.com*
**Corresponding author*

Abstract: *Micro-expression recognition plays a significant role in psychological analysis and public safety. To address the challenges of short duration and subtle motion interference in micro-expressions, this paper proposes an Asymmetric Dual-channel Fusion Network (ADF-Net) based on attention mechanisms. The method employs the Convolutional Block Attention Module (CBAM) to adaptively enhance the spatial structural features of static images, integrates Spatial Pyramid Pooling (SPP) to capture multi-scale motion details in dynamic images, and utilizes Squeeze-and-Excitation (SE) attention to achieve adaptive recalibration of dual-channel weights. Experimental results demonstrate that the proposed method achieves UF1 scores of 0.9062, 0.7718, and 0.7099 on the CASME II, SAMM, and SMIC benchmark datasets, with a composite UF1 score of 0.7513 on the 3DB-combined dataset.*

Keywords: *Micro-expression Recognition, Dual-channel Network, Dynamic Imaging, Asymmetric Pooling, Attention Mechanism, Feature Fusion*

1. Introduction

Micro-expression (ME) is a spontaneous and uncontrollable facial muscle movement with a short duration (1/25 to 1/5 second), making it difficult to suppress voluntarily [1]. Due to their extremely subtle motion amplitudes and complex spatiotemporal characteristics, achieving high-robustness micro-expression recognition (MER) has become an important research topic in affective computing.

Early MER methods primarily relied on handcrafted features such as Local Binary Patterns on Three Orthogonal Planes (LBP-TOP) [2] and optical flow histograms [3], which exhibit poor generalization under complex conditions. The construction of high-quality spontaneous databases—SMIC [4], CASME II [5], and SAMM [6]—has driven the transition to deep learning approaches. Various architectures have been explored: Van et al. [7] employed capsule networks to model facial spatial hierarchies; Zhang et al. [8] utilized graph structures to learn dynamic dependencies among facial regions; Yang et al. [9] leveraged contextual Transformers to capture long-range inter-frame dependencies; Miao et al. [10] combined Inception with CBAM for multi-scale representations. Verma et al. [11] proposed LEARNet, a dynamic imaging network for MER. ResNet [12] serves as a foundational backbone in many MER pipelines, including semi-lightweight multi-feature architectures [13] and cross-database compression strategies [14]. Dual-channel networks have emerged as the mainstream paradigm by processing static and dynamic information through parallel pathways: Wang et al. [15] embedded temporal convolution in a dual-stream framework; Wang et al. [16] introduced cross-layer attention and similarity constraints (DSN-CASC); Zhou et al. [17] exploited multi-scale Inception properties for cross-database MER. Zhao et al. [18] proposed a two-stage 3D CNN with optical flow as the core dynamic modality.

However, among existing approaches, advanced models often incur substantial computational overhead and are difficult to deploy. Moreover, conventional dual-channel architectures suffer from insufficient spatiotemporal decoupling: the two channels learn features in relative isolation, lacking effective interaction mechanisms. Additionally, many methods over-rely on optical flow as the sole source of dynamic information, which is computationally expensive and sensitive to illumination variations.

To address the above challenges, this paper proposes an Asymmetric Dual-channel Fusion Network (ADF-Net) based on attention mechanisms for micro-expression recognition. The method adopts an

ISSN 2616-5775 Vol. 9, Issue 6: 37-46, DOI: 10.25236/AJCIS.2026.090606

asymmetric feature extraction architecture: the RGB channel introduces CBAM[19] and Grid Pooling to enhance key spatial representations, while the dynamic channel utilizes Rank Pooling[20] combined with SPP[21] to capture multi-scale motion. SE-Attention[22] is introduced as a late-stage feature fusion module to achieve adaptive calibration of dual-channel weights.

2. Asymmetric Dual-channel Fusion Network

This paper proposes an Asymmetric Dual-Channel Fusion Network (ADF-Net) for micro-expression recognition. As shown in Figure 1, ADF-Net consists of four components: a static spatial channel with CBAM [19] and Grid Pooling, a dynamic temporal channel with Rank Pooling [20] and SPP [21], a dynamic image generation module, and an SE-Attention-based feature fusion module [22]. After dual-channel feature concatenation and adaptive recalibration, the network outputs three expression categories: Positive, Negative, and Surprise. Focal Loss with Label Smoothing is adopted to address class imbalance.

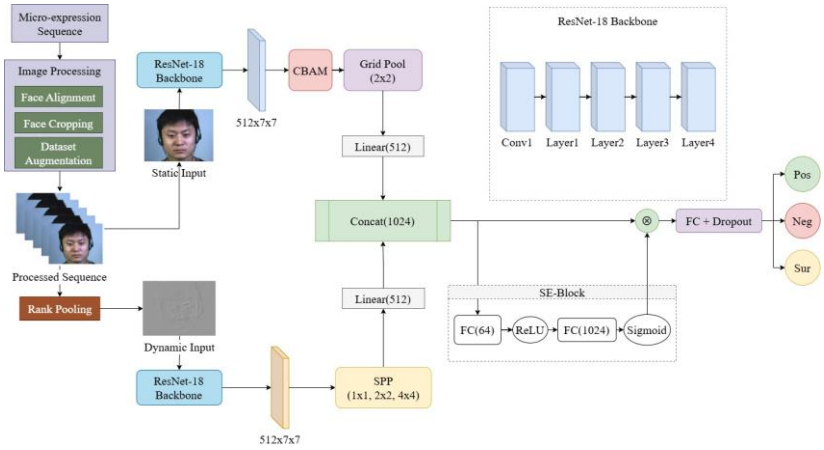


Figure 1. Asymmetric Dual-Channel Fusion Network (ADF-Net)

2.1 Static Channel Design

The static channel extracts facial spatial features. To avoid apex-frame annotation bias, a random frame is sampled from the active interval (onset to offset) for training. The channel uses ImageNet-pretrained ResNet-18 (GAP removed) as the backbone, producing a 7×7 feature map $F_s \in R^{512\times7\times7}$, which is then refined by CBAM [19] along both channel and spatial dimensions to emphasize discriminative regions such as the eyes and mouth corners, as shown in Figure 2.

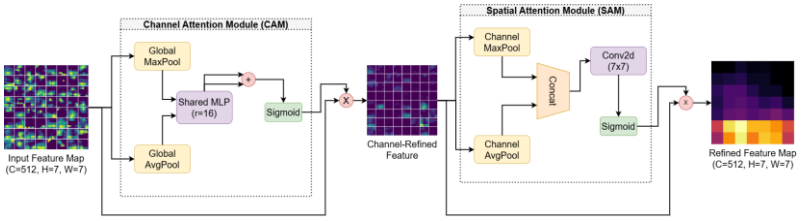


Figure 2. Structure of the Convolutional Block Attention Module (CBAM)

CBAM comprises a Channel Attention Module (CAM) and a Spatial Attention Module (SAM). CAM aggregates spatial context via global average and max pooling, then learns channel weights through a shared MLP, as expressed in Equation (1):

$$M_C(F_s) = \sigma(MLP(AvgPool(F_s)) + MLP(MaxPool(F_s))) \tag{1}$$

SAM pools along the channel axis and generates a spatial weight map through a 7×7 convolution to focus on discriminative locations, as expressed in Equation (2):

$$M_S(F_s') = \sigma(f^{7\times7}([AvgPool(F_s'); MaxPool(F_s')])) \tag{2}$$

The overall feature refinement process is expressed in Equation (3):

$$F_s'' = M_S(F_s') \otimes F_s' \tag{3}$$

ISSN 2616-5775 Vol. 9, Issue 6: 37-46, DOI: 10.25236/AJCIS.2026.090606

Through this dual attention mechanism, the feature map is adaptively weighted to suppress background and enhance key facial cues.

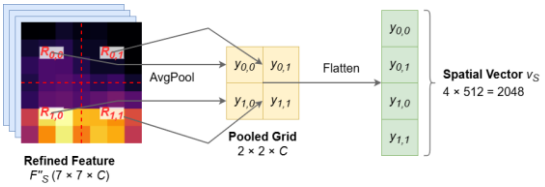


Figure 3. 2×2 Grid Pooling strategy

To prevent the severe spatial information loss caused by conventional GAP, which collapses a 7×7 map to 1×1, this paper adopts a 2×2 Grid Pooling strategy after CBAM, as shown in Figure 3.

Grid Pooling partitions the refined feature map into 4 spatial regions $R_{i,j}(i,j \in \{0,1\})$ and performs adaptive average pooling to each, yielding a 2×2×C representation, as expressed in Equation (4):

$$y_{i,j} = \frac{1}{|R_{i,j}|} \sum_{(h,w)\in R_{i,j}} F_s''(h,w) \tag{4}$$

The four region features are flattened and concatenated into a 4×C (2048-dim) vector, then projected to 512-dim for fusion. This preserves the spatial topology of facial expressions while achieving efficient feature aggregation.

2.2 Dynamic Channel Design

The dynamic channel captures subtle pixel-level motion from micro-expression sequences. It takes the Dynamic Image as input and uses ResNet-18 (GAP removed) as the backbone to produce feature maps $F_t \in R^{C\times H\times W}(C=512, H=7, W=7)$.

Given the diversity of motion amplitudes in micro-expressions, a single-scale pooling strategy is insufficient. A Spatial Pyramid Pooling (SPP) module [21] is therefore incorporated after the backbone, as shown in Figure 4, to capture multi-scale motion cues across varying receptive fields.

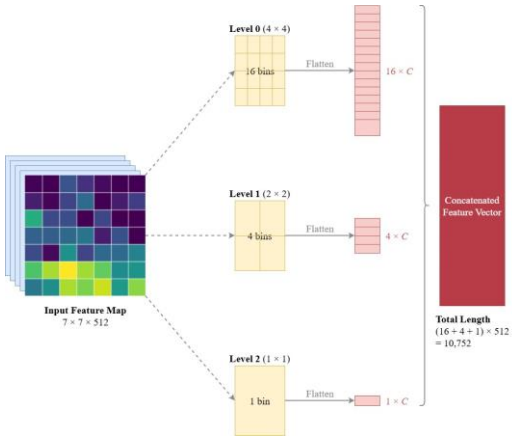


Figure 4. Spatial Pyramid Pooling module

In the SPP module, this paper defines three pooling levels $L=\{1,2,4\}$, with adaptive pooling operations performed at each level $l \in L$, as expressed in Equation (5):

$$v^{(l)} = Flatten(MaxPool_{l\times l}(F_t)) \tag{5}$$

The feature map is partitioned into grids of equal size (e.g., 4×4, 2×2, 1×1), and the maximum response of each cell is extracted to preserve key motion signals while reducing dimensionality.

The pooled features from different levels are concatenated to construct a multi-scale dynamic feature representation, as expressed in Equation (6):

$$v_t = Concat(v^{(1)}, v^{(2)}, v^{(4)}) = [v^{(1)}; v^{(2)}; v^{(4)}] \tag{6}$$

ISSN 2616-5775 Vol. 9, Issue 6: 37-46, DOI: 10.25236/AJCIS.2026.090606

The resulting 10752-dimensional vector is then projected to a 512-dim dynamic feature vector. This multi-scale design enables the model to simultaneously perceive local subtle movements (4×4 partitions) and global motion patterns (1×1), enhancing robustness across varying expression intensities.

2.3 Dynamic Image Generation

The Dynamic Image compresses the motion information of a video sequence into a single RGB image via Rank Pooling [20], preserving key temporal characteristics while reducing computational overhead. The overall process is illustrated in Figure 5.

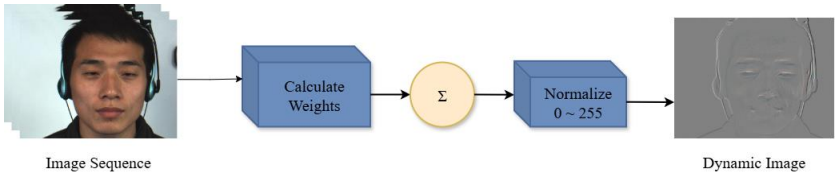


Figure 5. Dynamic Image Generation Process

The generation consists of two steps: temporal weight calculation and weighted aggregation. First, the temporal weight coefficient for each frame is computed according to Equation (7):

$$C_t = \sum_{j=t}^{N} \frac{2(j+1) - N - 1}{j+1} \tag{7}$$

Later frames receive higher cumulative weights, preserving motion saliency from onset to termination.

Once the weights are obtained, the final expression of the dynamic image is given by Equation (8):

$$D = N_{L1}\left(\sum_{t=1}^{N} C_t I_t\right) \tag{8}$$

The result is normalized to [0, 255] and converted to uint8 format.

2.4 Feature Fusion Strategy

The static feature vector and dynamic feature vector are first concatenated into a joint representation of dimension D=1024, as expressed in Equation (9):

$$U = Concat(v_s, v_t) \in R^{1024} \tag{9}$$

Since the contributions of static and dynamic features vary across samples—pronounced motion favors dynamic cues, subtle motion relies on static texture—this paper introduces the Squeeze-and-Excitation (SE) module [22] for adaptive recalibration, as shown in Figure 6.

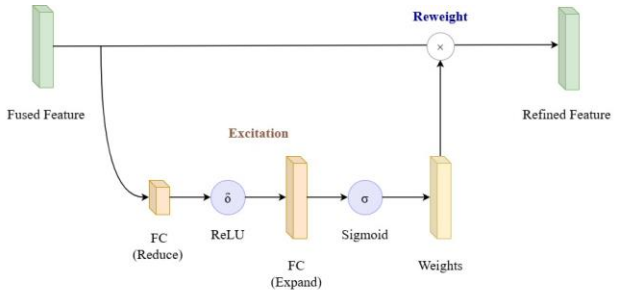


Figure 6. SE-based adaptive feature fusion module

(a) Squeeze-and-Excitation Stage: A dual fully-connected structure learns nonlinear interactions among features and generates normalized weight vectors, as expressed in Equation (10):

$$w = \sigma(W_2 \cdot \delta(W_1 \cdot U)) \tag{10}$$

The weights are normalized to (0, 1). Through channel-dimensional compression and expansion, the synergistic roles of different feature dimensions are explicitly modeled.

ISSN 2616-5775 Vol. 9, Issue 6: 37-46, DOI: 10.25236/AJCIS.2026.090606

(b) Adaptive Recalibration Stage: The weight vector is applied to the original features, achieving fine-grained recalibration, as expressed in Equation (11):

$$U_{refined} = w \odot U \tag{11}$$

This enables the network to adaptively balance the two channels: enhancing dynamic weights when motion is pronounced and strengthening static weights when motion is subtle. The refined feature vector is used for classification.

2.5 Loss Function

A modified Focal Loss combined with Label Smoothing is adopted for end-to-end multi-class optimization. Focal Loss dynamically down-weights easy samples via a modulating factor, focusing training on hard samples. The standard form is given in Equation (12):

$$L_{focal} = -(1 - p_{t,i})^{\gamma} log(p_{t,i}) \tag{12}$$

where $p_{t,i}$ is the model's predicted probability for the true class, and γ is the focusing parameter controlling the rate of weight decay. This paper sets $\gamma = 2.0$. When a sample is classified with high confidence, the modulating factor suppresses its loss contribution; for hard samples the weight approaches 1.

To prevent overfitting to training-set label distributions, Label Smoothing with factor is introduced, defining the smoothed target distribution in Equation (13):

$$q_c = \begin{cases} 1 - \alpha_{ls}, c = \text{target} \\ \dfrac{\alpha_{ls}}{K-1}, c \neq \text{target} \end{cases} \tag{13}$$

The final training objective combines the two, computing weighted cross-entropy between the predicted distribution and smoothed labels, as expressed in Equation (14):

$$L_{total} = \frac{1}{B} \sum_{i=1}^{B} [(1 - p_{t,i})^{\gamma} \sum_{c=1}^{K} (-q_{c,i} log(p_{c,i}))] \tag{14}$$

where B denotes the number of classes. This design alleviates the long-tail effect in micro-expression datasets and improves recognition stability under cross-subject (LOSO) evaluation.

3. Experiments

3.1 Dataset Preprocessing and Experimental Setup

Experiments are conducted on three benchmark datasets—CASME II [5], SAMM [6], and SMIC [4]—with a composite dataset (3DB-combined) constructed following MEGC2019 [23] for cross-database evaluation. Sample distributions are summarized in Table 1.

Table 1. Distribution of micro-expression samples

Category	CASME II	SAMM	SMIC	3DB-combined
Positive	32	51	26	109
Negative	99	70	92	261
Surprise	25	43	15	83
Total	156	164	133	453

All datasets undergo a standardized preprocessing pipeline:

(1) Face detection and alignment via MTCNN, cropping and aligning faces to 224×224 pixels with pixel normalization, as visualized in Figure 7;

(2) Static channel input constructed by randomly sampling one frame within the micro-expression active interval for implicit temporal augmentation;

(3) Online data augmentation including random horizontal flipping, rotation, color jitter, and affine transformation, as shown in Figure 8.

ISSN 2616-5775 Vol. 9, Issue 6: 37-46, DOI: 10.25236/AJCIS.2026.090606

Figure 7. Visualization of face images before and after alignment

Figure 8. Visualization of online data augmentation strategies

All models are trained end-to-end for 60 epochs with a batch size of 16. Each experiment is repeated three times, and the average is reported. Dataset-specific strategies are adopted: CASME II and SAMM use AdamW with Focal Loss and Label Smoothing; SMIC uses Adam with Focal Loss (no additional augmentation); 3DB-combined uses AdamW with Focal Loss. All settings include warm-up, cosine annealing, gradient clipping, and early stopping.

Evaluation follows the Leave-One-Subject-Out (LOSO) protocol. For N subjects, N folds are run, each holding out one subject for testing. The composite dataset follows MEGC2019 [23] procedures: category indices are first mapped, then merged, and cross-validated. Metrics include Accuracy (ACC), Unweighted F1 Score (UF1), and Unweighted Average Recall (UAR), with UF1 and UAR being more robust under class imbalance.

Experiments were run on Ubuntu 22.04 with an NVIDIA GPU (32 GB), Python 3.12, PyTorch 2.8.0, and CUDA 13.0.

3.2 Experimental Results

Figure 9 presents the training and validation accuracy curves of ADF-Net across all four datasets. Training accuracy converges within 20–30 epochs, while validation accuracy plateaus after 30–40 epochs without significant oscillations. A gap between training and validation curves persists, indicating mild overfitting—a common issue in MER given limited sample sizes and high inter-subject variability—but the absence of complete validation degradation suggests acceptable stability and generalization.

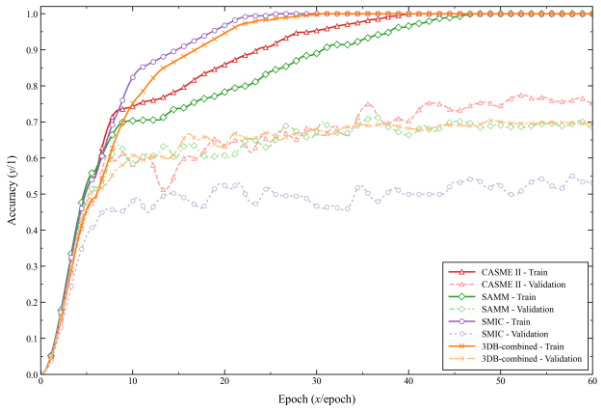


Figure 9. Training and validation accuracy curves of the ADF-Net across datasets

ISSN 2616-5775 Vol. 9, Issue 6: 37-46, DOI: 10.25236/AJCIS.2026.090606

3.2.1 Performance Comparison

To position ADF-Net within the MER landscape, we compare against traditional handcrafted methods and recent deep learning approaches, as shown in Table 2.

It should be noted that differences in evaluation protocols, category mapping, and preprocessing across published methods may affect comparability. This paper selects methods following the same LOSO protocol and comparable configurations to ensure fairness.

Deep learning methods consistently outperform handcrafted approaches. On CASME II, ADF-Net achieves 0.9062 UF1 and 0.8927 UAR, surpassing RCN and other lightweight models while matching more complex architectures such as DSN-CASC, confirming the effectiveness of asymmetric pooling in preserving spatial details. On SAMM, the UF1 of 0.7718 exceeds that of MERSiamC3D, validating dynamic feature stability under complex lighting.

On 3DB-combined and SMIC, ADF-Net lags behind top-tier methods such as MERSiamC3D. This stems from their differing feature extraction strategies: for SMIC—a high-frame-rate, low-resolution dataset—3D-CNN methods exploit dense spatiotemporal structure, whereas Rank Pooling compresses the temporal dimension into a single frame, which preserves holistic motion stability but may lose fine-grained motion details.

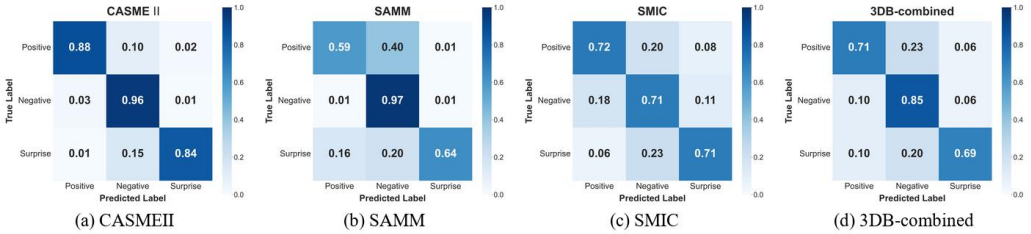
In summary, ADF-Net achieves competitive performance on most benchmarks, demonstrating strong generalization. However, in low-resolution or highly complex scenarios, the single-branch dynamic compression architecture still trails top-tier 3D-CNN models, yet offers an effective architectural reference for robust MER.

Table 2. Comparison of model performance

Method	3DB-combined		CASME II		SAMM		SMIC	
	UAR	UF1	UAR	UF1	UAR	UF1	UAR	UF1
LBP-TOP [2]	0.5882	0.5785	0.7429	0.7026	0.4102	0.3954	0.5280	0.2000
Bi-WOOF [3]	0.6227	0.6296	0.8026	0.7805	0.5139	0.5211	0.5829	0.5727
CapsuleNet [7]	0.6506	0.6520	0.7018	0.7068	0.5989	0.6209	0.5877	0.5820
OFVIG-Net [8]	0.6632	0.6720	0.7195	0.7129	0.5787	0.6066	0.6435	0.6400
RCN [14]	0.7190	0.7432	0.8778	0.8632	0.7418	0.7164	0.7256	0.7127
DFR-Net [13]	0.7295	0.7657	0.8491	0.8669	0.7325	0.7889	0.6579	0.6422
DSN-CASC [16]	0.7607	0.7516	0.8995	0.9074	0.6897	0.6782	0.6727	0.6633
MERSiamC3D [18]	0.7986	0.8068	0.8763	0.8818	0.7280	0.7475	0.7598	0.7356
CoTDPN [9]	0.7648	0.7424	0.7931	0.8015	0.7539	0.7367	0.6740	0.6321
DSNTC [15]	-	-	0.8921	0.8082	0.7876	0.7166	0.7531	0.6952
Dual-Inception [17]	0.7278	0.7353	0.8560	0.8621	0.6726	0.6645	0.5663	0.5868
Inception-CBAM+ [10]	0.7435	0.7420	0.8522	0.8562	0.7157	0.6931	0.6678	0.6693
ADF-Net (Ours)	0.7485	0.7513	0.8927	0.9062	0.7351	0.7718	0.7113	0.7099

3.2.2 Confusion Matrix Analysis

Figure 10 presents the confusion matrices of ADF-Net on each dataset.



(a) CASMEII (b) SAMM (c) SMIC (d) 3DB-combined

Figure 10. Confusion matrix on different datasets

From the confusion matrices, the following conclusions can be drawn:

ISSN 2616-5775 Vol. 9, Issue 6: 37-46, DOI: 10.25236/AJCIS.2026.090606

(1) On CASME II, the model achieves recall rates of 88%, 96%, and 84% for Positive, Negative, and Surprise, respectively, indicating robust discriminative capability despite pronounced class imbalance.

(2) In the 3DB-combined experiment, despite significant domain shift, recall rates reach 71% (Positive), 69% (Surprise), and 85% (Negative), suggesting that the dual-branch architecture and SE-Attention recalibration mitigate cross-domain performance degradation.

(3) On SMIC, recall rates for all three categories remain in the 71%–72% range, demonstrating that dynamic motion cues effectively complement static texture under low-resolution, high-noise conditions.

(4) On SAMM, approximately 40% of Positive and 20% of Surprise samples are misclassified as Negative, reflecting the common MER challenge that fine-grained expression cues are easily masked by inter-subject and illumination variability.

Overall, ADF-Net exhibits relatively balanced performance across datasets and partially alleviates class bias. While localized confusion persists in challenging scenarios such as SAMM, the results on CASME II and 3DB-combined indicate that the issue has not evolved into complete class collapse.

3.2.3 Ablation Study

To verify the contribution of each component, systematic ablation experiments are conducted on CASME II under LOSO cross-validation with ResNet-18 as the backbone. By progressively integrating SPP, CBAM, 2×2 Grid Pooling, and the SE module, results are reported in Tables 3 and 4.

Table 3. Ablation experiment-1 (Performance)

Model	Feature Extraction	Fusion Strategy	ACC	UAR	UF1
Base-S	ResNet18 (GAP)	/	0.6966	0.6496	0.6323
Base-D	ResNet18 (GAP)	/	0.8590	0.8612	0.8393
Model A	ResNet18 (GAP)	Direct Concat	0.8761	0.8463	0.8520
Model B	ResNet18 (SPP)	Direct Concat	0.9038	0.8679	0.8822
Model C	Grid+CBAM / SPP	Direct Concat	0.9103	0.8854	0.8908
ADF-Net	Grid+CBAM / SPP	SE Fusion	0.9252	0.8927	0.9062

Table 4. Ablation experiment-2 (Efficiency)

Model	Params/M	FLOPs/G	FPS	Latency/ms
Base-S	11.18	1.82	476.99	2.10
Base-D	11.18	1.82	439.89	2.27
Model A	22.36	3.65	226.37	4.42
Model B	33.37	3.66	205.80	4.86
Model C	28.94	3.65	197.51	5.06
ADF-Net	29.07	3.65	214.61	4.66

From the analysis of the ablation experiment results, the following conclusions can be drawn:

(1) Core status and effectiveness of dynamic information: Baseline-D (85.90% ACC) outperforms Baseline-S (69.66% ACC) by 16.24 percentage points, despite both having identical parameter counts and computational costs (11.18M / 1.82G FLOPs). This demonstrates that the dynamic image generated by Rank Pooling preserves crucial discriminative temporal evolution information at a low computational cost. Dynamic information, compared with static textures, holds irreplaceable core value in MER tasks.

(2) Effectiveness of dual-channel architecture and the dilemma of concatenation-based fusion: After directly concatenating static and dynamic features, Model A achieves 87.61% accuracy, but with doubled parameter count (22.36M). However, its UAR (0.8463) is lower than that of Baseline-D (0.8617), indicating that direct concatenation introduces feature redundancy and that spatial and motion information interfere with each other. This validates the need for refined fusion mechanisms.

(3) SPP multi-scale pooling is a key contributor to performance improvement: Replacing the terminal GAP of the dynamic branch with SPP in Model B raises accuracy from 87.61% to 90.38% (+2.77 percentage points) and UF1 from 0.8731 to 0.8951. This represents the largest single-module improvement in the ablation experiments. The multi-scale receptive field provided by SPP enables the model to simultaneously capture local subtle movements and global motion patterns, which is critical for micro-expression scenarios characterized by diverse motion amplitudes.

(4) Rationality of the asymmetric pooling design: Model C replaces SPP in the static branch with CBAM + 2×2 Grid Pooling, reducing parameters from 33.37M to 28.94M (a reduction of approximately

ISSN 2616-5775 Vol. 9, Issue 6: 37-46, DOI: 10.25236/AJCIS.2026.090606

4.4M) while further improving accuracy to 91.03% and UF1 to 0.9024. This result validates the core design rationale of this paper—namely, applying differentiated, asymmetric pooling strategies according to the representational characteristics of static and dynamic information.

(5) High efficiency, light overhead, and significant gains of SE fusion. The final ADF-Net adds only 0.13M parameters over Model C (29.07M vs. 28.94M) with FLOPs unchanged (3.65G), yet UF1 improves from 0.9024 to 0.9062 and ACC reaches 92.52%. SE-Attention effectively mitigates the mutual interference between dual-channel features through adaptive recalibration. With negligible computational overhead (+0.45% parameters), it delivers a 1.49% UF1 improvement, demonstrating an extremely high efficiency-to-performance ratio.

It is also important to note that the FLOPs and inference latency statistics in Table 4 only account for the forward propagation stage. In practical deployment, the Rank Pooling-based dynamic image generation processes input sequences significantly faster than traditional optical flow computation. Baseline-D achieves 85.90% accuracy with only 11.18M parameters, further corroborating the high efficiency of Rank Pooling as a lightweight temporal modeling approach. Therefore, when considering the entire inference pipeline, ADF-Net achieves an even more favorable balance between effectiveness and efficiency compared with typical dual-branch approaches.

4. Discussion

This paper proposes an asymmetric dual-channel fusion network, ADF-Net, for micro-expression recognition. In contrast to conventional symmetric dual-branch architectures, ADF-Net designs differentiated feature extraction strategies tailored to the inherent differences in representation between static texture information and dynamic motion information. In the static channel, CBAM and Grid Pooling are employed to adaptively enhance critical spatial features while preserving fine-grained spatial structural information. In the dynamic channel, dynamic images are combined with Spatial Pyramid Pooling to enhance the modeling of multi-scale motion variations in micro-expressions. At the fusion stage, SE-Attention is further introduced to adaptively calibrate dual-channel features, alleviating the noise introduction and contribution imbalance caused by simple concatenation.

Experimental results demonstrate that ADF-Net achieves highly competitive recognition performance on multiple benchmark datasets, with improvements across several evaluation metrics compared with existing methods, indicating that the proposed asymmetric architecture and adaptive fusion strategy are effective. At the same time, the experiments also reveal that under conditions of small sample sizes, high inter-subject variability, and extremely low expression intensity, the model still has room for further improvement. Future research will focus on stronger generalization constraints, finer-grained local feature modeling, and the extension to integrated detection-recognition pipelines for real-world scenarios.

Acknowledgements

This work was supported by the Joint Research and Development Program of Shanghai Shenkang Hospital Development Center and United Imaging (2022SKLY 12).

References

[1] Ekman P, Friesen W V. Nonverbal leakage and clues to deception[J]. Psychiatry, 1969, 32(1): 88-106.
[2] Zhao G, Pietikainen M. Dynamic texture recognition using local binary patterns with an application to facial expressions[J]. IEEE Trans on pattern analysis and machine intelligence, 2007, 29(6): 915-928.
[3] Liong S T, See J, Wong K S, et al. Less is more: Micro-expression recognition from video using apex frame[J]. Signal Processing: Image Communication, 2018, 62: 82-92.
[4] Li X, Pfister T, Huang X, et al. A spontaneous micro-expression database: Inducement, collection and baseline [C]// The 10th IEEE International Conference and Workshops on Automatic face and gesture recognition (fg). Piscataway, NJ: IEEE Press, IEEE, 2013: 1-6.
[5] Yan W J, Li X, Wang S J, et al. CASME II: An improved spontaneous micro-expression database and the baseline evaluation[J]. PLoS One, 2014, 9(1): e86041.
[6] Davison A K, Lansley C, Costen N, et al. Samm: A spontaneous micro-facial movement dataset[J]. IEEE Trans on affective computing, 2016, 9(1): 116-129.

ISSN 2616-5775 Vol. 9, Issue 6: 37-46, DOI: 10.25236/AJCIS.2026.090606

[7] Van Q N, Chun J, Tokuyama T. CapsuleNet for micro-expression recognition [C]// The 14th IEEE international conference on automatic face & gesture recognition (FG 2019). Piscataway, NJ: IEEE Press, IEEE, 2019: 1-7.
[8] Zhang L, Zhang Y, Sun X, et al. Micro-expression recognition based on direct learning of graph structure[J]. Neurocomputing, 2025, 619: 129135.
[9] Yang J, Wu Z, Wu R. Micro-expression recognition based on contextual transformer networks[J]. The Visual Computer, 2025, 41(3): 1527-1541.
[10] Miao Yuhan, Huang Shucheng. Micro-expression recognition based on Inception-CBAM+ [J]. Computer Applications and Software, 2026, 43(1): 171-177.
[11] Verma M, Vipparthi S K, Singh G, et al. LEARNet: Dynamic imaging network for micro expression recognition[J]. IEEE Trans on Image Processing, 2019, 29: 1618-1627.
[12] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition [C]// Proceedings of the IEEE conference on computer vision and pattern recognition. Piscataway, NJ: IEEE Press, 2016: 770-778.
[13] Li M, Huang X, Wu L. A Semi-Lightweight Multi-Feature Integration Architecture for Micro-Expression Recognition[J]. Computers, Materials & Continua, 2025, 84(1).
[14] Xia Z, Peng W, Khor H Q, et al. Revealing the invisible with model and data shrinking for composite-database micro-expression recognition[J]. IEEE Trans on Image Processing, 2020, 29: 8590-8605.
[15] Wang H, Wang K, Yu W. The dual stream network with embedding temporal convolution for micro-expression recognition[J]. Signal, Image and Video Processing, 2025, 19(5): 403.
[16] Wang G, Huang S. Dual-stream network with cross-layer attention and similarity constraint for micro-expression recognition[J]. Multimedia Systems, 2024, 30(3): 147.
[17] Zhou L, Mao Q, Xue L. Dual-inception network for cross-database micro-expression recognition [C]// The 14th IEEE international conference on automatic face & gesture recognition (FG 2019). Piscataway, NJ: IEEE Press, IEEE, 2019: 1-5.
[18] Zhao S, Tao H, Zhang Y, et al. A two-stage 3D CNN based learning method for spontaneous micro-expression recognition[J]. Neurocomputing, 2021, 448: 276-289.
[19] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module [C]// Proceedings of the European conference on computer vision (ECCV). Piscataway, NJ: IEEE Press, 2018: 3-19.
[20] Fernando B, Gavves E, Oramas J, et al. Rank pooling for action recognition[J]. IEEE Trans on pattern analysis and machine intelligence, 2016, 39(4): 773-787.
[21] He K, Zhang X, Ren S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Trans on pattern analysis and machine intelligence, 2015, 37(9): 1904-1916.
[22] Hu J, Shen L, Sun G. Squeeze-and-excitation networks [C]// Proceedings of the IEEE conference on computer vision and pattern recognition. Piscataway, NJ: IEEE Press, 2018: 7132-7141.
[23] See J, Yap M H, Li J, et al. MEGC 2019—the second facial micro-expressions grand challenge [C]// The 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019). Piscataway, NJ: IEEE Press, IEEE, 2019: 1-5.