

Article

PAST: Prior-Aware Sparse Transformer for Micro-Expression Recognition

Jiateng Liu ^{1,2} , Tianchen Zhou ^{3,*}, Hengcan Shi ^{4,*}, Yining Zhao ⁴, Zedong Liu ⁵ , Yingtian Yu ⁵ and Liming Liu ⁵ 

¹ Key Laboratory of Child Development and Learning Science (Southeast University), Ministry of Education, Nanjing 211102, China; jiateng_liu@seu.edu.cn

² School of Biological Science and Medical Engineering, Southeast University, Nanjing 211102, China

³ School of Computer Science and Engineering, Southeast University, Nanjing 211102, China

⁴ School of Artificial Intelligence and Robotics, Hunan University, Changsha 410082, China; fzdlpzt@hnu.edu.cn

⁵ China Mobile Communications Group Hunan Co., Ltd., Changsha 410005, China; liuzedong@hn.chinamobile.com (Z.L.); yuyintian@hn.chinamobile.com (Y.Y.); liuliming2@hn.chinamobile.com (L.L.)

* Correspondence: ztc0526@126.com (T.Z.); shihengcan@hnu.edu.cn (H.S.)

Abstract

Micro-expression recognition (MER) has a lot of applications in lie detection, education, healthcare, etc., as involuntary micro-expressions (MEs) may provide subtle facial cues associated with affective responses. With the development of deep learning, many studies have recently employed Vision Transformers (ViTs) to investigate MER, since ViTs show promising performance in various visual domains due to their excellent local–global modeling ability. However, such methods confront two fundamental challenges: First, fine-grained visual features are needed to capture the subtle facial movements of MEs, which ViTs relatively fall short on due to coarse patch resolution constrained by their quadratic complexity. Second, the data-intensive nature of ViTs impedes effective learning given the limited scale of ME data. To overcome the aforementioned limitations of using ViTs for MER, we propose the Prior-aware Sparse Transformer (PAST), a novel Transformer-based architecture integrating spatial and semantic prior knowledge synergistically into a sparse attention mechanism, enabling linear-complexity processing of large amounts of fine-grained features. Specifically, we first designed an extraction algorithm to generate a representative set of motion-intensive Principal Anchors, which are used to guide the model's focus on biologically critical regions during sampling. Second, we introduced the Semantic Dictionary, which was trained with a carefully designed self-contrastive loss to embed task-invariant discriminative semantics of the anchors. Such global semantics further modulate patch sampling and attention weighting in the sparse attention procedure, achieving better training performance with limited ME data. Extensive evaluations on MEGC and CD6ME protocols demonstrate state-of-the-art performance, validating PAST's efficacy for MER.

Keywords: micro-expression recognition; knowledge-guided learning; sparse attention; self-supervised contrastive learning



Academic Editor: George A. Tsihrintzis

Received: 9 June 2026

Revised: 19 July 2026

Accepted: 21 July 2026

Published: 28 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Micro-expressions (MEs) involve subtle facial muscle movements that occur in highly localized regions. Because MEs may contain subtle facial cues related to affective responses, micro-expression recognition (MER) is practical in many applications, such as lie detection, medical diagnosis, and education, which have become research hotspots in the area of

artificial intelligence and affective computing [1–5]. However, MER is still challenging due to the subtle motions of ME, which are hard to capture, and it is necessary to present a powerful approach to handle MER.

With the development of computer vision and deep learning, deep models demonstrate powerful abilities in various tasks. Convolutional Neural Networks (CNNs) were the first to be applied in MER, thanks to their ability of local signal modeling [6–8]. Recently, Vision Transformers (ViTs) [9] have achieved promising success in computer vision due to their excellence at capturing long-range dependencies and show great performance in various domains. However, despite some successful attempts of ViTs regarding MER [10–12], their performances remain suboptimal for MER compared with other vision tasks.

The main reason for this can be divided into two views due to the localized and sparse nature of ME signals. On the one hand, vanilla ViT attention cannot effectively capture the localized motions of MEs, since typical self-attention functions incur quadratic computational complexity, posing limitations to the input scale. ViTs practically employ large patch sizes (e.g., 14×14 , 16×16 , or 32×32) [13,14]. While these are suitable for most vision tasks, such a coarse patch configuration critically constrains the model performance during MER, blurring the localized motion signals that are essential for recognizing MEs and inducing noise within patches. On the other hand, the need for a large amount of training data makes it difficult to train a strong ViT model for MER. The global attention function computes weights between all patch pairs, most of which are irrelevant to MER due to the sparsity of informative patches. This introduces redundancy that impedes convergence and optimization during training [15,16]. The limited scale of ME datasets further hinders a model's ability to learn relevant global attention patterns. To address these limitations, we propose a method whose overall pipeline is illustrated in Figure 1.

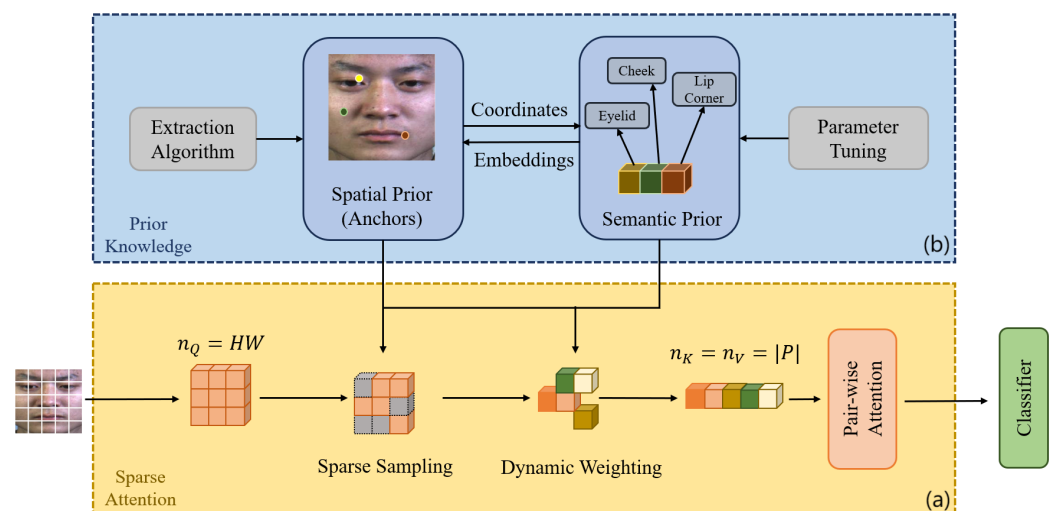


Figure 1. Pipeline of our method. Multiple-source prior knowledge guides the filtering and weighting of the sparse pair-wise attention, enabling (a) higher-resolution visual patches within an affordable computational load and (b) effective parameter learning on data-limited MER.

Sparse attention addresses such constraints by selectively reducing the amount of attention pairs, enabling finer-grained feature extraction. Such methods require excessive hyperparameter tuning for the best performance. Deformable attention [17], inspired by deformable convolutions [18], dynamically samples informative regions with reference point grids. While enhancing ViT's flexibility and efficiency, this uniform sampling strategy lacks task-specific priors, and thus, is inefficient for the highly structured and spatially sparse nature of ME data.

To achieve an effective sparse attention representation with an affordable computational cost, we propose the Prior-aware Sparse Transformer (PAST), which incorporates spatial and semantic prior knowledge into Transformer attention. Our sparse attention architecture is based on the core concept of **Principal Anchors**, a set of keypoints derived from the 68 facial landmarks based on local motion intensity. These anchors serve as the initialization of sampling locations. Instead of using uniform grids [17], we impose spatial priors by starting from motion-intensive facial regions. Taking advantage of the sparse sampling pipeline, the attention function remains computationally affordable, even with extremely large amounts of fine-grained patches. Moreover, as the anchors provide consistent semantic contexts, semantic priors can be utilized to further enhance training and inference. Therefore, we introduce the **Semantic Dictionary**, a global embedding matrix optimized by a self-supervised contrastive loss to embed high-level semantics of each respective anchor. Task-invariant global semantics inject high-level facial structures into the model. By integrating semantic priors into the adaptive token sampling and dynamic attention weighting processes, we address the difficulties of training Transformer-based deep models under an ME data shortage. Our contributions can be summarized as follows:

1. For motion sparsity, we employed a biologically anchored sparse attention mechanism integrated with **spatial priors** to enable the processing of large finer-grained patches without losing critical information, therefore capturing localized ME signals. We designed an extraction algorithm to generate Principal Anchors from facial landmarks based on motion intensity, the coordinates of which serve as the initial sampling locations to suppress irrelevant noise and highlight important regions.
2. To stabilize the training of our Transformer-based model with limited ME data and further improve its performance, we proposed refining the process of sparse attention with **semantic priors**, injecting informative semantics into the model via the Semantic Dictionary, a global semantic embedding matrix that encodes discriminative semantic representations of each Principal Anchor. We designed a self-contrastive training paradigm to impose facial structures and symmetry patterns into the learning of the Semantic Dictionary.
3. Extensive experiments were conducted to demonstrate PAST's efficacy, achieving state-of-the-art performance on several benchmark MER datasets and protocols.

2. Related Work

2.1. Deep Learning-Based Micro-Expression Recognition

Convolutional Neural Networks (CNNs) were used in the earliest successful attempts at learning-based MER. OFF-ApexNet [6] utilizes horizontal and vertical directions of optical flow through dual CNN streams for classification. Zhou et al. [19] employed Graph Convolutional Networks (GCNs) to model the relationship of Action Units (AUs) as an enhancement to label classification. Li et al. [20] proposed the Spatial and Channel Attention (SCA) weighting module for AU-guided attention learning. Addressing mislabeled samples, Varanka et al. [21] developed an iterative label correction method to prevent overfitting and error accumulation. Wang et al. [22] tackled dataset scale limitations using adversarial learning for sample interpolation, enhancing temporal feature modeling through Inception networks and 3DCNNs. Ruan et al. [23] introduced a region replacement module for data augmentation using ResNet18 [24] as the backbone, which was optimized with an AU similarity-based contrastive loss. While these CNN-based methods improve local feature modeling, their limited receptive field restricts joint modeling of distant facial regions relevant to a single ME. With the development of Transformer models [14], Transformer-based approaches have advanced micro-expression recognition. Zhai et al. [12] extracted multi-scale features and fused them via Transformer attention. Fan et al. [11]

proposed a ViT-based model [9] using contrastive learning for symmetric facial region representation. Yu et al. [25] leveraged self-supervised pretraining with VideoMAE [26], freezing pretrained parameters and training only lightweight adapters for downstream micro-expression recognition.

However, these Transformer-based methods inherit vanilla ViT's coarse patch resolution and data-intensive training requirements, neither of which is addressed by explicit facial priors. In this study, we aimed to tackle these problems with spatial-prior-aware sparse attention. The biologically anchored sparse sampling can effectively filter significant patches from among large-scale candidates resulting from fine-grained patch settings, allowing for the modeling of localized motion signals without adding a notable computational cost. Moreover, the need for excessive training data is alleviated to some extent, since irrelevant patches are explicitly suppressed from entering the attention function instead of requiring the model to implicitly learn the modeling of such redundancy.

2.2. Efficient Attention in Transformer

To achieve more efficient attention computation and easier optimization, previous research employed specific sparse patterns in Transformer models. The original self-attention mechanism has a quadratic complexity regarding the input length, making it computationally expensive for tasks involving large sets of tokens and difficult to converge with limited data. Sparse attention trims unnecessary computations by limiting the calculations to a subset of token pairs, focusing on the most relevant relations while ignoring less important ones, significantly reducing the computational costs while maintaining or even improving the performance. Zaheer et al. [27] proposed BigBird, a Turing complete approximator of sequence functions that combines local, global, and random attention to achieve linear complexity. Reformer [16] employs locality-sensitive hashing to approximate nearest neighbors. Therefore, they proposed projecting keys and values into a lower-dimensional space to achieve efficient attention. DAT [17] introduced a dynamic sampling mechanism by attending to a grid of reference points for each query, rather than fixed patterns, enabling more flexible and adaptive attention computation while retaining high efficiency. However, this uniform sampling grid is agnostic to task-specific structure, making it inefficient for ME data where informative motion concentrates in sparse, anatomically consistent regions. Sparse attention serves as a vital tool for efficiently handling long and noisy sequences in various domains.

In summary, existing sparse attention designs rely on fixed or uniformly initialized sampling, without exploiting structured facial priors—a gap that is especially costly given the scarcity of ME training data. Motivated by this, we present our model, PAST, which integrates prior knowledge into the sparse attention pipeline to make Transformers adapt to the MER domain better.

Beyond MER-specific attention design, injecting external priors to guide visual representation learning has also proven effective in broader multimodal settings. For instance, ViLoMem [28] employs a dual-stream memory to separately encode visual and logical priors for continual multimodal reasoning, while VEGA-3D [29] repurposes implicit 3D priors learned by video generation models to enhance MLLMs' spatial understanding. These works further support the effectiveness of explicitly modeling structured prior knowledge to compensate for the data- or supervision-scarcity challenges faced by large vision models, echoing the motivation behind our prior-aware sparse attention design.

2.3. Fairness in Facial Analysis

Beyond accuracy and efficiency, an increasing body of work highlights that facial analysis systems can exhibit systematic performance disparities across demographic subgroups.

Buolamwini and Gebru [30] conducted an intersectional audit of commercial gender classification systems and revealed substantially higher error rates on darker-skinned and female faces compared with lighter-skinned and male faces, underscoring the risks of deploying facial analysis systems without subgroup-level evaluation. To address such disparities, Datta and Swamidass [31] proposed Fair-Net, a branched multitask architecture that reduces classification and calibration gaps across identifiable sub-populations in class-imbalanced datasets through architectural modifications to the output layer and loss function. These works motivate the need to consider demographic fairness when deploying facial-analysis-based models such as PAST in sensitive applications.

3. Method

To enable effective modeling of subtle ME movements, high visual granularity of visual input is necessary, which leads to small patch sizes and, in turn, produces an excessively large number of tokens. A sparse attention mechanism is therefore essential to sample a small number of important tokens that preserve critical information. To achieve this goal, we propose the Prior-aware Sparse Transformer (PAST), which incorporates facial prior knowledge to guide token sampling and attention weighting. The overall architecture of our model is introduced in Section 3.1.

3.1. Overview

As illustrated in Figure 2, our model employs a hierarchical structure with a downsampling block between layers. Specifically, we adopted 4×4 patches to acquire the initial token sequence, yielding a 56×56 initial feature map from the 224×224 input of optical flow. Practically, we implement patch embedding with a learnable convolutional layer, with a kernel size of 4×4 and a stride of 4. In this study, we utilized the optical flow of 3 channels: the horizontal component u , the vertical component v , and the magnitude $\|f\|$.

Our architecture consists of L attention layers. In each layer L_i , the sizes of the feature maps remain consistent and are passed through N_i attention blocks, e.g., vanilla attention, local attention, and our proposed PAST. Each layer ends with a downsampling block that downsamples the feature map $\mathbf{X}_i \in \mathbb{R}^{B \times H \times W \times d}$ to half of its size, which becomes the input for the next layer $\mathbf{X}_{i+1} \in \mathbb{R}^{B \times \frac{H}{2} \times \frac{W}{2} \times 2d}$. The knowledge database is shared across all layers. Finally, a classifier is imposed on the final feature map to output the classification probabilities of N classes.

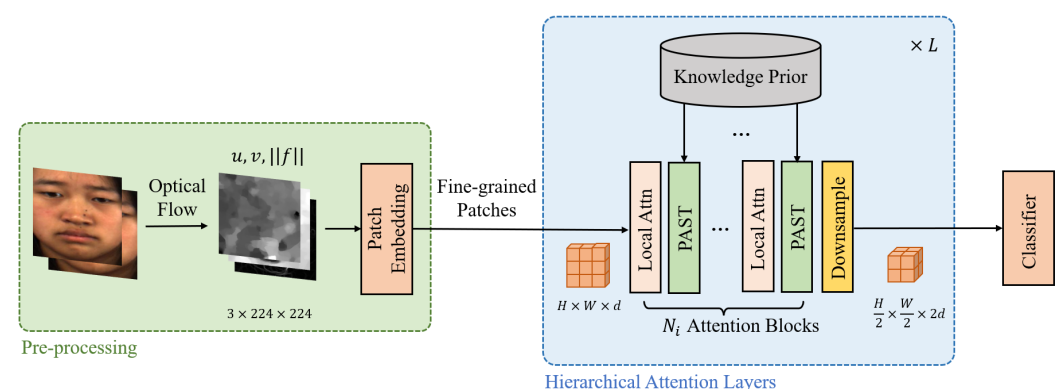


Figure 2. The hierarchical structure of our model. Each layer consists of several attention blocks, with multiple knowledge priors injected into the training and inference. Optical flow features are downsampled after each layer to achieve a multi-scale representation. Finally, a classifier is applied to output prediction probabilities.

3.2. Principal Anchors

The sparse attention mechanism can meet the computational constraints, even with excessively large fine-grained input. However, without proper prior guidance, models cannot effectively learn the sparse representation given the localized ME signals. The incentive of Principal Anchors lies in the fact that muscle movements in certain facial regions are highly correlated with MEs. Therefore, we aimed to explicitly filter such important regions and remove irrelevant noise through Principal Anchors.

We generated Principal Anchors from the 68 facial landmarks, directly or combinationally, as they are a precise indicator of specific facial regions and can be accurately extracted with existing toolkits. To select N_p optimal Principal Anchors that best cover facial movement signals, we needed to quantify the enrichment of motion signals at different positions. Therefore, we first introduced a scalar metric termed **Facial Movement Coverage**, i.e., C , that measures the intensity of motion at given areas based on the optical flow distribution of the samples within the dataset. For a point set P , $C(P)$ is defined as

$$C(P) = \frac{1}{|S|} \sum_{s \in S} \sum_{(x,y) \in \bigcup_{p \in P} \mathcal{N}^k(p)} \sqrt{\mathbf{u}_{s,x,y}^2 + \mathbf{v}_{s,x,y}^2}, \quad (1)$$

where $\mathcal{N}^k(p) = \{(m,n) \in \mathbb{Z}^2 \mid |m - x_p| \leq k, |n - y_p| \leq k\}$ denotes a $(2k+1) \times (2k+1)$ neighborhood centered at $p = (x_p, y_p)$, $\mathbf{u}_{x,y}$ and $\mathbf{v}_{x,y}$ are the horizontal and vertical components of optical flow at (x,y) , and S is a subset of the dataset. $C(P)$ quantifies motion intensity with the average magnitude within the union neighborhood area $\bigcup_{p \in P} \mathcal{N}^k(p)$ of all points in P . In this study, we used the 68 facial landmarks L as the candidate point set and approximated the global location of each facial landmark with its average coordinates across the subset S . Although the average coordinates may not perfectly align with the intended anatomical locations on every subject in the dataset, this approximation is justified as the difference is negligible after our preprocessing pipeline involving face cropping, resizing, and alignment.

Then, we employed a greedy selection algorithm to iteratively expand the Principal Anchor set I until $|I| \geq N_p$, whose process is shown in Algorithm 1. We manually selected 3 pairs of symmetric landmarks as the initial anchor set $I_D = I_{DL} \cup I_{DR} = \{20, 39, 60\} \cup \{25, 44, 56\}$. At each iteration, Algorithm 1 selects a single landmark p_l that maximizes the coverage gain $A(p, I) = C(I \cup \{p\}) - C(I)$ from the left hemiface or the central axis, and then adds both p_l and its symmetric counterpart p_r into I to ensure the symmetric structure in the output. If the candidate is on the central axis, the landmark alone will be added into I .

For specific regions around the cheeks that may contain frequent movement but lack description landmarks, we augmented I with six combinational anchors $\{c_1, \dots, c_6\}$ following [32]:

$$I \leftarrow I \cup \{c_1, c_2, c_3, c_4, c_5, c_6\}, \quad (2)$$

where each anchor c_i is derived from linear combinations of existing landmarks:

$$\begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \\ c_5 \\ c_6 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} p_{43} \\ p_{43} \\ p_{43} \\ p_{40} \\ p_{40} \\ p_{40} \end{bmatrix} - \frac{1}{2} \begin{bmatrix} p_{40} \\ p_{40} \\ p_{40} \\ p_{43} \\ p_{43} \\ p_{43} \end{bmatrix} + \begin{bmatrix} p_{30} \\ p_{36} \\ p_{55} \\ p_{30} \\ p_{32} \\ p_{49} \end{bmatrix}, \quad (3)$$

which yields a set of $N_p + 6$ Principal Anchors as the final anchor set.

Algorithm 1 Principal Anchor Extraction**Require:** Initial anchor set I_D , landmark set L , target anchor count N_p **Ensure:** Principal Anchor set I

```

1:  $I \leftarrow I_D$  // Initialize Principal Anchors
2:  $R \leftarrow L \setminus I_D$  // Initialize set of candidate landmarks
3: while  $|I| < N_p$  do
4:   for each  $p \in R$  do
5:      $A(p, I) \leftarrow C(I \cup \{p\}) - C(I)$  // Compute coverage increment for each candidate anchor on the
        left-hemiface or central axis
6:   end for
7:    $p_l \leftarrow \arg \max_{p \in R} A(p, I)$ 
8:   if  $p_l$  is on the left-hemiface then
9:      $p_r \leftarrow$  symmetric landmark of  $p_l$  // Add the symmetric counterpart if existing
10:     $I \leftarrow I \cup \{p_l, p_r\}$ 
11:     $R \leftarrow R \setminus \{p_l, p_r\}$ 
12:   else
13:     $I \leftarrow I \cup \{p_l\}$ 
14:     $R \leftarrow R \setminus \{p_l\}$ 
15:   end if
16: end while
17: return  $I$ 

```

3.3. Semantic Dictionary

Since Principal Anchors are derived directly or combinationally from the facial landmarks, each anchor corresponds to an anatomically consistent facial location. Therefore, anchors carry discriminative high-level semantics, which is beneficial for MER. For example, an eyelid raise may indicate surprise or curiosity, while a lip corner raise may represent more positive emotions. To leverage these input-independent semantics, we integrate a learnable feature matrix $\mathbf{D} \in \mathbb{R}^{|I| \times d_{\text{dict}}}$ into the model, termed the **Semantic Dictionary**, where each element $\mathbf{D}_i$ encodes the aforementioned semantic representation of the i -th Principal Anchor as a d -dimensional feature vector. As a global embedding matrix, $\mathbf{D}$ remains unchanged for all samples in the same pass. $\mathbf{D}$ injects semantic information into the process of token sampling and attention weighting, which is detailed in Section 3.4.

During training, $\mathbf{D}$ is optimized jointly with other parameters. However, the absence of direct supervision signals may lead to unstable training and model degradation. For example, feature vectors may collapse to similar representations, losing the ability to represent discriminative semantics. To ensure effective learning of distinctive embeddings, we introduce a self-supervised contrastive loss on $\mathbf{D}$:

$$\mathcal{L}_{\text{dict}} = -\frac{1}{|I|} \sum_{i=1}^{|I|} \log \frac{\exp(1/\tau) + \exp(\mathbf{z}_i \cdot \mathbf{z}_{i^+} / \tau)}{\sum_{j=1}^{|I|} \exp(\mathbf{z}_i \cdot \mathbf{z}_j / \tau)}, \quad (4)$$

where $\mathbf{z}_i = \frac{\mathbf{D}_i}{\|\mathbf{D}_i\|_2}$ denotes the ℓ_2 -normalized embedding vector, i^+ indexes the symmetric counterpart of anchor i , and τ is the temperature factor. $\mathcal{L}_{\text{dict}}$ enforces inter-anchor discriminability in the embedding space based on facial symmetry. For bilaterally symmetric anchors, their embedding vectors $\mathbf{z}_i$ and $\mathbf{z}_{i^+}$ form positive pairs, promoting similarity in the contrastive process, since symmetric regions may share similar high-level semantics. For anchors on the central axis, $\mathbf{z}_{i^+}$ is defined as $\mathbf{0}$, simplifying the numerator to $\exp(1/\tau)$ to enforce any difference between each other. Finally, all other pairs serve as negative pairs. Thus, $\mathcal{L}_{\text{dict}}$ encourages inter-anchor discriminability while preserving similarities for symmetric anchors. The final loss function is defined as

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda \mathcal{L}_{\text{dict}}, \quad (5)$$

where λ is a balancing factor.

3.4. Facial-Knowledge-Guided Sparse Attention

Existing methods utilize a uniform sampling grid to achieve efficient attention [17], which lacks representative capabilities for ME data where relevant movements are highly localized and structured. For the localization problem, we integrated a spatial prior into the original pipeline to reduce redundancy and encourage more focus on critical spots. For the structural facial input, we achieved semantics-aware inference with the guidance of a semantic prior.

Our proposed sparse attention module PAST introduces prior knowledge into the model with two essential steps. First, reference points are initialized with the coordinates of Principal Anchors. Second, the Semantic Dictionary provides global contexts to modulate the offset generation and attention weight computation. This architecture is illustrated in Figure 3.

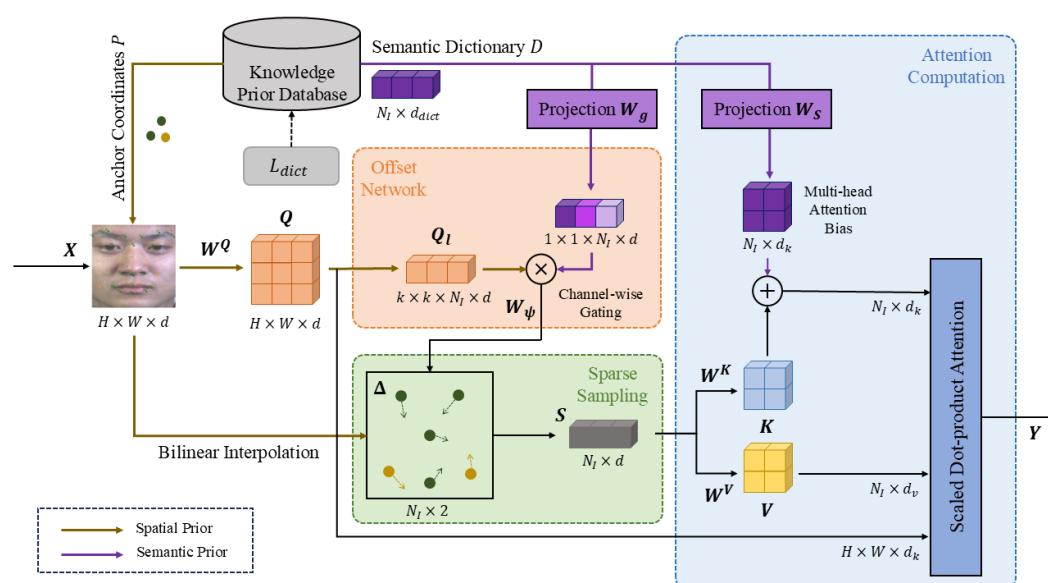


Figure 3. The architecture of the proposed sparse attention mechanism in PAST. The information flow of two types of prior knowledge is denoted with arrows in different colors. Principal Anchors, whose coordinates are adjusted by a lightweight offset network, are utilized for sparse sampling. The sampling process is implemented with bilinear interpolation to handle continuous coordinates. Semantic priors, described by the Semantic Dictionary, are employed in the whole process. First, they serve as a gating signal in the offset network to refine offset generation and sampling locations. Second, they act as a bias term in the attention-weighting process to dynamically adjust the weights with the awareness of specific facial regions.

3.4.1. Offset Network

The offset network $\psi : \mathbf{X} \mapsto \Delta \in \mathbb{R}^{B \times |I| \times 2}$ generates 2D offsets for each Principal Anchor $p = (x_p, y_p)$. The token at $(x_p + \Delta x_p, y_p + \Delta y_p)$ is then sampled into the sparse sequence. Thus, a single anchor will generate 1 sampled token in the sequence, yielding $|I|$ tokens in total.

To effectively inject prior knowledge while keeping the offset network lightweight, we incorporated the Semantic Dictionary into it as a gating signal. Given the query tensor $\mathbf{Q} \in \mathbb{R}^{B \times H \times W \times d}$, we first extracted local features $\mathbf{Q}_I \in \mathbb{R}^{B \times k \times k \times |I| \times d}$ via $k \times k$ cropping around each anchor. The Semantic Dictionary D gates visual features on the channel dimension:

$$g(\mathbf{D}) = \sigma(\mathbf{D}\mathbf{W}_g) \in \mathbb{R}^{1 \times 1 \times |I| \times d}, \quad (6)$$

where σ is the sigmoid function and $\mathbf{W}_g$ is a learnable projection matrix. Offsets are then computed using

$$\Delta = \mathbf{W}_\psi(g(\mathbf{D}) \otimes \mathbf{Q}_I) \in \mathbb{R}^{B \times |I| \times 2}, \quad (7)$$

where $\otimes$ denotes broadcasted element-wise multiplication. $\mathbf{W}_\psi$ is practically implemented with two $k \times k$ convolutional kernels. Finally, offsets are cropped within the range $[-\frac{k}{2}, \frac{k}{2}]$, with the field receptive by $\mathbf{Q}_I$.

3.4.2. Sparse Token Sampling and Attention Weight Computation

After obtaining sampling coordinates $P' = P + \Delta$, we sampled tokens via interpolation:

$$\mathbf{S} = \phi(\mathbf{Q} \mid P') \in \mathbb{R}^{B \times |I| \times d}, \quad (8)$$

where ϕ denotes bilinear interpolation, since the sampling operation is non-differentiable on discrete indices, and the offsets output from the offset network are inherently continuous.

Finally, we employed a multi-head scaled dot-product attention to acquire attention weights and outputs. Sparse key-value pairs are computed using

$$\begin{cases} \mathbf{K} = \mathbf{S}\mathbf{W}^K + \mathbf{D}\mathbf{W}_S & \in \mathbb{R}^{B \times |I| \times d_k} \\ \mathbf{V} = \mathbf{S}\mathbf{W}^V & \in \mathbb{R}^{B \times |I| \times d_v}, \end{cases} \quad (9)$$

where $\mathbf{W}^K$ and $\mathbf{W}^V$ are learnable projections, and d_k and d_v denote the feature dimensions for keys and values, respectively, in a single attention head. The projection layer $\mathbf{W}_S(D)$ fuses semantic information as a bias term when acquiring keys. Specifically, each token $\mathbf{S}_i$ generated by anchor i is fused with its corresponding embedding vector $\mathbf{D}_i$, enabling facial semantics to dynamically modulate attention for respective anchors. Keys and values are token sequences of length $|I|$, leading to a subsequent computation complexity of $\mathcal{O}(|I| \cdot HW)$ for the scaled dot-product attention, which is significantly reduced from $\mathcal{O}(H^2W^2)$, since $|I| \ll HW$ holds for small patches.

This design of sparse attention ensures that the model focuses on motion-intensive regions only, effectively suppressing irrelevancy to maintain computational efficiency. Given limited ME data, this advantage of PAST further makes convergence easier during training. The semantic prior is fused through the whole process, thereby enabling the model's awareness of facial structures and enhancing its ability to dynamically adjust its focus regarding different areas.

4. Experiments

4.1. Evaluation Protocols and Datasets

We evaluated our model on two cross-dataset evaluation protocols: MEGC and CD6ME, comprising seven public datasets in total. Both protocols are detailed in the following sections.

4.1.1. MEGC

The Micro-Expression Grand Challenge (MEGC) [33] combines three benchmark datasets: CASME II [34], SAMM [35], and SMIC [36]. MEGC employs the Leave-One-Subject-Out (LOSO) cross-validation across 68 subjects, classifying expressions into three categories: Positive, Negative, and Surprise. Model performance is measured using the unweighted F1-score and unweighted average recall:

$$\text{UF1} = \frac{1}{C} \sum_{i=1}^C \frac{2TP_i}{2TP_i + FN_i + FP_i}, \quad \text{UAR} = \frac{1}{C} \sum_{i=1}^C \frac{TP_i}{N_i}, \quad (10)$$

with the final metrics averaged over $C = 3$ classes. LOSO prevents subject-specific overfitting caused by data imbalance across subjects. However, it requires training 68 splits for a complete round of evaluation.

4.1.2. CD6ME

To address historical problems in MEGC including data unreliability and data leakage, CD6ME [37] incorporates six datasets (CASME [38], CASME II [34], CAS(ME)³-A [39], SAMM [35], MMEW [40], 4DME [41]) using Leave-One-Dataset-Out (LODO) cross-validation, resulting in six splits for complete evaluation. Due to label inconsistencies across datasets, CD6ME reformulates MER as a multi-label AU detection task for 12 frequent AUs (1, 2, 4, 5, 6, 7, 9, 10, 12, 14, 15, 17). CD6ME uses UF1 across AUs as the evaluation metric. It is worth noting that CD6ME explicitly prohibits early stopping to prevent data leakage.

4.2. Implementation Details

We utilized PyTorch version 2.1.0 to implement our model. We employed TV-L1 algorithm to obtain optical flow maps for all samples, with consistent face cropping and resizing. Face detection and alignment were performed using dlib's frontal face detector combined with a five-point similarity transform, aligning eye centers to fixed reference coordinates prior to cropping to 224×224 ; samples where dlib failed to detect a face or landmarks were excluded from the training and evaluation. Optical flow was computed between the onset and apex frames with TV-L1 parameters $\tau = 0.25$, $\lambda = 0.15$, and $\theta = 0.3$, and five pyramid levels. To extract Principal Anchors, we used the dlib toolkit for accurate landmark detection. Five hundred random samples from CAS(ME)³-C [39] were used as set S for Algorithm 1, with a neighborhood radius $k = 4$ pixels. Note that we used CAS(ME)³-C to capture priors, while our testing dataset, CD6ME, included CAS(ME)³-A. There is no overlap between them. We extracted $N_p = 18$ landmarks with indices 18, 20, 22, 23, 25, 27, 28, 32, 34, 36, 39, 44, 49, 52, 55, 56, 58, and 60, with six combinational anchors (described in Section 3.2) as Principal Anchors.

For the model hierarchy, we employed $L = 4$ layers in our model architecture, with $N_i = [2, 4, 10, 2]$ attention blocks in each layer. Then, 4×4 patches were extracted from optical flow maps of size 224×224 as the model's input. The Semantic Dictionary dimension was set to 64, with $N_s = 8$ attention heads used throughout all attention modules, and the local crop size k in the offset network was set to 7×7 . For Layers 1 and 2, we adopted the same attention blocks as in DAT. For Layer 3, we alternated between local attention [42] and our proposed PAST to construct this layer. Finally, Layer 4 consisted of $N_4 = 2$ PAST blocks before the MLP classifier.

During training, a stochastic gradient descent optimizer (momentum 0.9, weight decay 5×10^{-4}) was employed with an initial learning rate of 7×10^{-5} , which gradually declined to 1×10^{-5} using a cosine annealing scheduler over 150 epochs, with a batch size of 16 and 10 warm-up epochs. Data augmentation included random horizontal flipping. We employed $\mathcal{L}_{\text{cls}}$ as the cross-entropy loss, with weights inversely proportional to the number of samples for each class. The balancing factor λ was set to 0.2. The temperature factor in $\mathcal{L}_{\text{dict}}$ was set to 0.5. All experiments were conducted on four NVIDIA RTX 4090 GPUs. Unless otherwise noted, all reported results were averaged over three independent runs with random seeds {0, 1, 2}, and we observed a standard deviation below 0.5 UF1 points across runs.

4.3. Comparison with the State-of-the-Art

Table 1 presents our model's evaluation results on MEGC compared with other state-of-the-art methods. The accuracy and F1-score are reported using a scale of 0 to 100. As demonstrated in Table 1, our method maintains state-of-the-art performance on MEGC.

Notably, PAST achieves the highest UAR across all datasets, with an overall accuracy of 87.0%, demonstrating superior performance in general. Meanwhile, PAST achieves the second-highest UF1 on the full 3DB evaluation. Overall, there are eight results in Table 1, and we achieved the best performance on five out of them, showcasing PAST's robustness across multiple datasets.

Table 1. Comparison with the state of the art on MEGC.

Method	SMIC		CASME II		SAMM		3DB	
	UF1	UAR	UF1	UAR	UF1	UAR	UF1	UAR
LBP-TOP [43]	51.2	52.8	70.3	74.3	39.5	41.0	58.8	57.9
GEME [44]	60.4	63.9	88.3	87.9	58.4	54.6	72.2	73.0
MERSiamC3D [45]	73.6	76.0	88.2	87.6	77.5	72.8	80.7	79.9
SLSTT [46]	74.0	72.0	90.1	88.5	71.5	64.3	81.6	79.0
FRL-DGT [12]	74.3	74.9	91.9	90.3	77.2	75.8	81.2	81.1
TFT [47]	74.1	71.8	90.7	90.9	70.9	65.6	81.4	80.1
LAENet [48]	66.2	65.2	91.0	91.2	68.1	66.2	75.7	74.1
CoTDPN [49]	63.2	67.4	80.2	79.3	73.7	75.4	74.2	76.5
SDGSA [50]	<u>79.6</u>	<u>79.3</u>	95.0	94.1	<u>81.7</u>	80.9	85.4	<u>84.9</u>
HTNet [51]	80.5	79.1	95.3	<u>95.2</u>	81.3	<u>81.2</u>	86.0	84.6
Ours	77.6	79.5	<u>95.1</u>	95.5	84.6	85.0	<u>85.5</u>	87.0

Note: Bold numerical values indicate the best performance, and underlined numerical values indicate the second-best performance.

The confusion matrices for PAST on MEGC are shown in Figure 4, revealing an in-depth understanding into the model's per-class performance. Generally, PAST exhibits strong robustness across three datasets on samples labeled "Surprise" and "Positive". While PAST tends toward false positives for "Negative" instances in CASME II and SAMM, which may have originated from data imbalance within composite data, the error rates are relatively lower than other available methods. Nevertheless, the state-of-the-art performance shown in the combined results demonstrates PAST's effectiveness regarding MER.

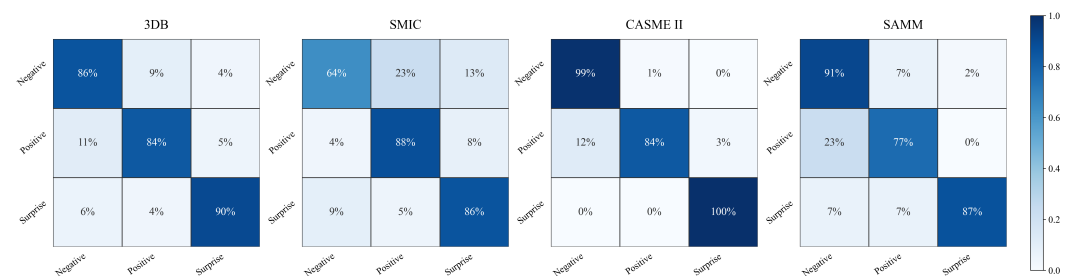


Figure 4. Confusion matrices for 3-class classification on 3DB, SMIC, CASME II, and SAMM.

Table 2 illustrates the performance of PAST on CD6ME. As a newer benchmark with fewer comparable works, a limited number of reports are available for comprehensive comparison. Therefore, we compared our method with only a few models with corresponding reports or accessible code. As suggested from Table 2, PAST achieves top-2 performances on nine of the 12 AUs, five of which with the highest UF1 results. PAST achieves the highest overall UF1 on CD6ME, with an improvement of 1.5 points over the second-highest method. Moreover, our standard deviation when using different datasets is 11.06 under UF1 and the CD6ME protocol; Micromamba's and SSSNet's standard deviations are 23.02 and 26.3, which are larger than ours. The results on two benchmarks with composite data demonstrates PAST's robustness and efficacy compared with other state-of-the-art methods.

Table 2. Comparison with the state of the art on CD6ME.

Method	AU1	AU2	AU4	AU5	AU6	AU7	AU9	AU10	AU12	AU14	AU15	AU17	UF1
LBP-TOP [43]	41.6	36.7	62.0	0.0	0.0	0.0	1.7	0.0	0.0	3.5	0.0	0.0	12.1
MDMO [52]	59.4	57.0	72.2	16.7	2.3	18.1	10.9	0.0	27.4	25.5	13.0	27.8	27.5
NMER [53]	19.1	19.2	43.3	9.3	6.3	9.1	12.5	4.8	18.1	22.6	3.3	6.4	14.5
SCA [20]	42.3	42.8	56.2	14.8	1.2	23.4	13.4	3.0	21.8	37.1	4.0	22.9	23.6
RCN-A [54]	74.8	71.0	85.3	4.7	0.0	24.1	15.9	0.0	21.1	23.7	0.0	24.0	28.7
RN18(2+1)D [55]	54.2	49.4	72.7	26.9	5.2	20.4	11.7	12.4	23.3	25.5	9.7	41.9	29.4
STSTNet [7]	74.2	68.8	85.7	3.6	6.4	29.0	18.6	9.0	16.7	29.0	11.8	28.3	31.8
ResNet18 [24]	<u>76.2</u>	<u>72.1</u>	<u>87.1</u>	12.6	5.5	38.2	8.6	18.4	37.3	33.7	12.4	44.7	37.2
Off-ApexNet [6]	74.9	70.2	86.3	13.5	3.3	<u>44.5</u>	<u>36.6</u>	18.3	32.0	37.7	19.0	38.2	39.5
SSSNet [21]	74.6	<u>72.1</u>	87.5	13.6	5.3	48.6	20.3	19.3	36.2	40.8	22.7	44.7	40.5
MicroMamba [56]	75.4	67.8	85.0	28.3	19.1	33.8	37.7	<u>21.2</u>	24.0	48.2	18.8	<u>50.4</u>	<u>42.5</u>
Ours	77.5	76.5	83.6	28.3	<u>10.3</u>	34.6	35.9	26.6	<u>36.4</u>	<u>45.9</u>	<u>19.8</u>	53.1	44.0

Note: Bold numerical values indicate the best performance, and underlined numerical values indicate the second-best performance.

4.4. Ablation Study

4.4.1. Model Efficiency and Performance with Different Attention Modules

We evaluated the computational efficiency of three attention modules using three metrics, i.e., FLOPs, memory usage, and inference time, using progressively finer patch sizes. The results are shown in Figure 5, with red crosses indicating Out-of-Memory (OOM) errors on a single GPU with 24 GB memory.

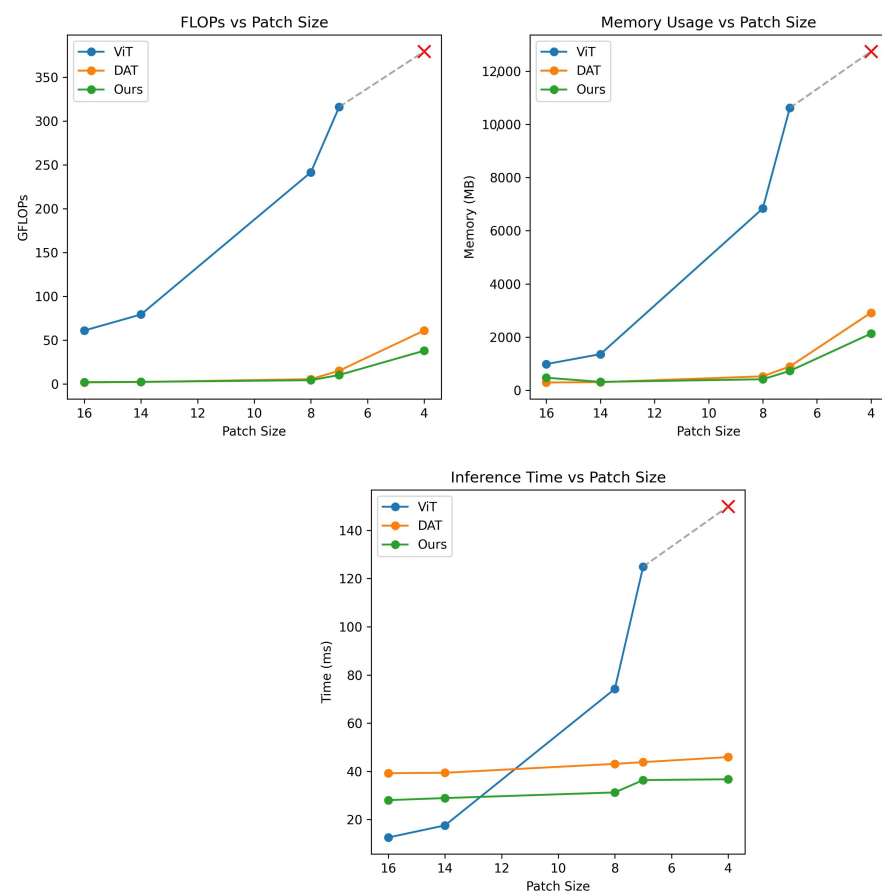


Figure 5. Computational efficiencies of attention modules at different patch sizes. Red crosses indicate Out-of-Memory (OOM) errors on a single GPU with 24 GB memory.

For self-attention [9], the computational cost grows exponentially with its quadratic complexity $\mathcal{O}(n^2)$. At patch size 8, FLOPs reach ~ 250 G. The training becomes impractical at patch size 4, with a theoretical ~ 991 G FLOPs for a batch size of 16, triggering OOM on 24 GB GPUs. All subfigures demonstrate that ViT's quadratic complexity poses limitation for fine-grained visual patches. Deformable attention [17] utilizes a uniformly distributed anchor grid that samples one token every K rows and columns, resulting in sub-quadratic complexity $\mathcal{O}(n^2/K^2)$. The sparse sampling enables finer patches, but the anchor count still scales with finer patches. Our sampling scheme with fixed cardinality maintains stable costs $\mathcal{O}(|I| \cdot n)$ across coarse-to-fine granularities. Despite extra parameters introduced by prior knowledge, PAST achieves significantly lower computational costs than other methods at patch sizes ≤ 8 .

We compared the three attention modules with different patch sizes on CD6ME (Table 3); the results suggest multiple findings. Comparing rows 1 and 2, we find that deformable attention alone provides no notable performance gain over ViT for coarse patch sizes. We argue that the significant loss of relevant visual information with coarse granularity is the direct reason for the relatively low performance. Contrasting deformable attentions for different patch sizes, we find that reducing patches to 4×4 improves UF1 by 2.1 points, confirming the necessity of fine-grained patches for capturing subtle MEs. Finally, our method achieves a +2.8 UF1 gain at identical fine-grained granularity, demonstrating facial priors' efficacy in filtering irrelevant features for data-limited MER.

Table 3. Ablation study of attention modules on CD6ME.

Method	Patch Size	AU1	AU2	AU4	AU5	AU6	AU7	AU9	AU10	AU12	AU14	AU15	AU17	UF1
Self-Attn [9]	16	71.7	65.6	78.5	26.3	19.4	30.3	34.3	22.4	25.4	45.2	8.7	38.9	38.9
Deformable Attn [17]	16	73.3	63.9	84.0	22.1	15.0	30.5	34.3	26.5	26.0	41.6	18.4	34.2	39.1
Deformable Attn [17]	4	74.3	71.8	83.5	26.6	17.8	30.7	31.5	25.6	27.8	45.4	19.5	40.2	41.2
Ours	4	77.5	76.5	83.6	28.2	10.3	34.6	35.9	26.6	36.4	45.9	19.8	53.1	44.0

Note: Bold numerical values indicate the best performance.

4.4.2. Effectiveness of Semantic Dictionary

We further performed an ablation study on the Semantic Dictionary, whose results are shown in Table 4. First, we removed both priors entirely, and the model is effectively reduced to vanilla deformable attention. Then, we involved the Principal Anchors only without the Semantic Dictionary. Subsequent rows evaluate different training strategies for the Semantic Dictionary. "None" in the "Loss" column denotes $\mathcal{L}_{\text{dict}} = 0$ throughout training. "Symmetry-agnostic" denotes our proposed contrastive loss with $\mathbf{z}_{i+} = \mathbf{0}$ for all anchors, i.e.,

$$\mathcal{L}_{\text{dict}} = -\frac{1}{|I|} \sum_{i=1}^{|I|} \log \frac{\exp(1/\tau)}{\sum_{j=1}^{|I|} \exp(\mathbf{z}_i \cdot \mathbf{z}_j / \tau)}, \quad (11)$$

discarding symmetry structures between anchor pairs. "Symmetry-aware" denotes the full loss-preserving anatomical symmetry patterns.

The results in Table 4 demonstrate various findings. First, the spatial prior embedded within Principal Anchors substantially enhances the model performance, as validated from the 1.1 points improvement, confirming the necessity of spatial guidance. Second, the Semantic Dictionary alone does not provide significant performance gains. However, when paired with appropriate self-supervision, it produces significant improvements, validated by the 0.9 points of UF1 increase when the symmetry-agnostic loss function is applied. Moreover, the full symmetry-aware loss achieves the highest performance score by 0.6,

demonstrating that the model’s capability regarding MER benefits greatly from proper supervision of semantic structures.

Table 4. Ablation study of semantic dictionary on CD6ME.

Anchors	Dictionary	Loss	UF1
✗	✗	N/A	41.2
✓	✗	N/A	42.3
✓	✓	None	42.5
✓	✓	Symmetry-agnostic	43.4
✓	✓	Symmetry-aware	44.0

Note: Bold numerical values indicate the best performance ✓ indicates that the component is used, while ✗ indicates that it is not used. N/A indicates not applicable

4.4.3. Effectiveness of the Dictionary Use Type

Table 5 disentangles the Semantic Dictionary’s two functional pathways on MEGC. The offset gating pathway alone (only OG) produces a small UAR improvement but a slight UF1 drop over the Principal Anchor-only baseline (only PA), while the key bias pathway (only KB) delivers substantial gains in UF1 and UAR. Their full integration, i.e., the whole PAST, achieves the best overall performance with a large margin, validating the strong complementarity of the two semantic injection mechanisms.

Table 5. Ablation results of dictionary use on MEGC.

Exp.	UF1	UAR
Only PA	77.5	73.4
Only OG	75.0	74.2
Only KB	78.6	79.7
Whole PAST	85.5	87.0

Note: Bold numerical values indicate the best performance.

4.4.4. Effectiveness of Key Components

Table 6 presents the results of component-wise ablation, including the patch size, sampling base, learned offsets, and dictionary use, with all experiments sharing an identical network hierarchical structure, training schedules, and data preprocessing pipelines to isolate the independent contribution of each module on MEGC under UF1 and UAR metrics. Specifically, first, comparing Exps. 1 and 2 confirms that fine-grained 4×4 patches deliver a better performance over the coarse 16×16 patches, verifying that high-resolution visual inputs are critical for capturing subtle localized ME motions. Subsequently, Exps. 2, 3, 5, and 6 disentangle the spatial priors of PAST. Motion-selected Principal Anchors consistently outperform uniform grids (i.e., the vanilla DAT), standard uniform facial landmarks, and randomly initialized reference points, demonstrating the efficiency and effectiveness of motion-intensity-guided anchor selection for the MER task in our PAST.

Furthermore, comparing Exps. 4 and 6 validates that learnable offsets can bring improvement by enabling adaptive sampling alignment for individual ME facial variations. In the end, Exps. 6 to 7 quantify the independent gain of the proposed semantic dictionary, fully validating the rationality of each key design in PAST.

Table 6. Component-wise controlled ablation results of PAST on MEGC.

Exp.	Patch Size	Sampling Base	Offsets	Dictionary	UF1	UAR
1	16 × 16	Uniform grid	✓	✗	57.3	57.9
2	4 × 4	Uniform grid	✓	✗	73.4	70.6
3	4 × 4	Uniform landmarks	✓	✗	74.2	70.8
4	4 × 4	Motion-selected	✗	✗	68.9	65.5
5	4 × 4	Random points	✓	✗	78.7	79.1
6	4 × 4	Principal Anchors	✓	✗	81.4	80.2
7	4 × 4	Principal Anchors	✓	✓	85.5	87.0

Note: Bold numerical values indicate the best performance ✓ indicates that the component is used, while ✗ indicates that it is not used.

4.5. Visualization

To qualitatively analyze the model discriminability across different label categories, we utilized t-Distributed Stochastic Neighbor Embedding (t-SNE), a non-linear dimensionality reduction technique to visualize the clustering of features in the embedding space [57]. We compared the features extracted using ViT, DAT, and our proposed PAST with three label classes. The visualization results are shown in Figure 6. Due to computational limitations discussed in Section 4.4, we adopted patch sizes of 16 for ViT and 4 for DAT and PAST. We find that PAST achieves similar levels of intra-class cohesion as DAT, and the features extracted by PAST show the most clear inter-class decision borders among the three attention mechanisms. ViT falls short on both metrics due to using coarse patches. This visualization further proves the discriminative ability of our model with fine-grained patches.

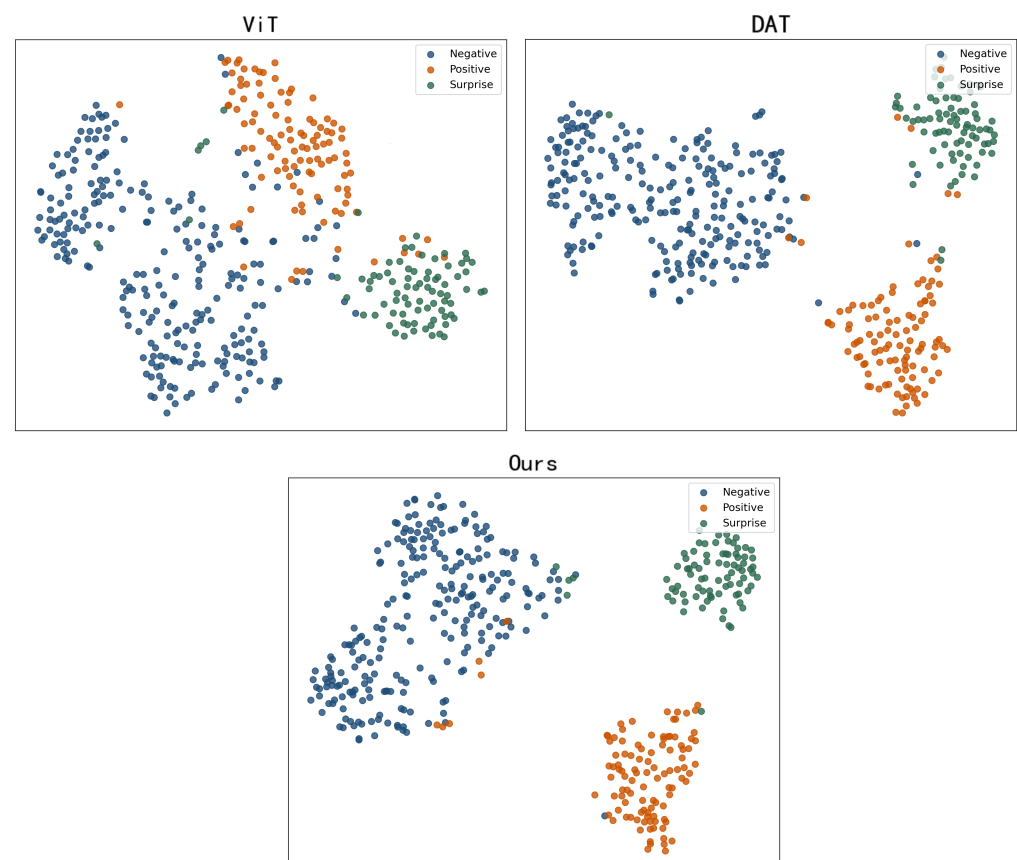


Figure 6. The visualization of feature clustering using t-SNE. Our method shows the most significant inter-class discriminability and intra-class cohesion across the three attention mechanisms.

5. Conclusions

In this study, we addressed the challenges of Transformer-based models in MER by proposing PAST, a novel attention mechanism that combines facial prior knowledge with sparse attention through three key innovations. First, Principal Anchors derived from motion-intensive facial areas provide spatially optimized reference points for sparse sampling, significantly reducing the computational complexity without loss of informative facial spots. Second, the Semantic Dictionary trained with a self-supervised contrastive loss encodes global facial semantics, enhancing the model's awareness of facial structure and symmetry patterns. Finally, our sparse attention mechanism dynamically fuses both types of aforementioned priors during inference, enabling efficient fine-grained processing of ME data.

PAST enables effective attention on fine-grained visual patches as a result of its linear complexity. With the guidance of spatial priors, PAST is capable of precisely capturing subtle localized ME signals. Moreover, the embedded semantics provide rich context regarding facial structure and further improve PAST's performance with limited training data. Extensive experiments and ablation studies demonstrate PAST's effectiveness. It is competitive on the MER task and outperforms other state-of-the-art methods on most metrics.

While these results demonstrate the effectiveness of PAST under standard MER protocols, the current evaluation is limited by the lack of complete demographic metadata in existing benchmarks, which prevents systematic subgroup-level fairness analysis. Future work should assess whether MER systems perform consistently across demographic groups and acquisition conditions before considering deployment in sensitive domains.

Author Contributions: Conceptualization, J.L. and T.Z.; methodology, J.L.; software, J.L. and T.Z.; validation, Y.Z., Z.L., Y.Y. and L.L.; formal analysis, J.L.; investigation, J.L. and T.Z.; resources, Z.L., Y.Y. and L.L.; writing—original draft preparation, J.L.; writing—review and editing, Y.Z., T.Z. and H.S.; visualization, J.L.; supervision, H.S.; project administration, H.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported in part by the National Natural Science Foundation of China under Grant 62401202, and in part by the large-model-based audio and vision emotion dataset project by the Hunan University–China Mobile Communications Group Co., Ltd. Industrial Intelligent Joint Institute, China Mobile Hunan Company Limited and China Mobile Communications Group Co., Ltd.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used in this study (CASME, CASME II, CAS(ME)³, SAMM, SMIC, MMEW, 4DME) are publicly available benchmark datasets and can be obtained from their respective providers. Dataset: CASME, accessible link: <http://casme.psych.ac.cn/casme/e1>, accessed date: 28 November 2024; dataset: CASME II, accessible link: <http://casme.psych.ac.cn/casme/e2>, accessed date: 5 November 2024; dataset: CAS(ME)³, accessible link: <http://casme.psych.ac.cn/casme/e4>, accessed date: 24 November 2024; dataset: SAMM, accessible link: <https://helward.mmu.ac.uk/STAFF/M.Yap/dataset.php>, accessed date: 17 December 2024; dataset: SMIC, accessible link: <https://www oulu.fi/en/university/faculties-and-units/faculty-information-technology-and-electrical-engineering/center-for-machine-vision-and-signal-analysis>, accessed date: 30 December 2024; dataset: MMEW, accessible link: <https://github.com/benxianyeteam/MMEW-Dataset>, accessed date: 12 January 2025; dataset: 4DME, accessible link: <https://www oulu.fi/en/university/faculties-and-units/faculty-information-technology-and-electrical-engineering/center-for-machine-vision-and-signal-analysis>, accessed data: 21 January 2025.

Conflicts of Interest: Author Mr. Zedong Liu, Mr. Yingtian Yu and Mr. Liming Liu were employed by the company the Hunan University–China Mobile Communications Group Co., Ltd. China Mobile Hunan Company Limited and China Mobile Communications Group Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

MER	Micro-expression recognition
ME	Micro-expression
ViT	Vision Transformer
CNN	Convolutional Neural Network
PAST	Prior-aware Sparse Transformer
AU	Action Unit
LOSO	Leave-One-Subject-Out
LODO	Leave-One-Dataset-Out
UF1	Unweighted F1-score
UAR	Unweighted average recall

References

1. Dong, Z.; Wang, G.; Lu, S.; Dai, L.; Huang, S.; Liu, Y. Intentional-deception detection based on facial muscle movements in an interactive social context. *Pattern Recognit. Lett.* **2022**, *164*, 30–39. [\[CrossRef\]](#)
2. Li, X.; Yi, X.; Ye, J.; Zheng, Y.; Wang, Q. SFTNet: A microexpression-based method for depression detection. *Comput. Methods Programs Biomed.* **2024**, *243*, 107923. [\[CrossRef\]](#) [\[PubMed\]](#)
3. Mahendar, M.; Malik, A.; Batra, I. Facial Micro-expression Modelling-Based Student Learning Rate Evaluation Using VGG–CNN Transfer Learning Model. *SN Comput. Sci.* **2024**, *5*, 204. [\[CrossRef\]](#)
4. Liu, J.; Shi, H.; Wang, Y.; Zong, Y. RF-REN: RGB-Frequency Relation Exploration Network for Micro-Expression Recognition. *IEEE Signal Process. Lett.* **2025**, *32*, 4439–4443. [\[CrossRef\]](#)
5. Zhao, S.; Tang, H.; Liu, S.; Zhang, Y.; Wang, H.; Xu, T.; Chen, E.; Guan, C. ME-PLAN: A deep prototypical learning with local attention network for dynamic micro-expression recognition. *Neural Netw.* **2022**, *153*, 427–443. [\[CrossRef\]](#) [\[PubMed\]](#)
6. Gan, Y.S.; Liong, S.T.; Yau, W.C.; Huang, Y.C.; Tan, L.K. OFF-ApexNet on micro-expression recognition system. *Signal Process. Image Commun.* **2019**, *74*, 129–139. [\[CrossRef\]](#)
7. Liong, S.T.; Gan, Y.S.; See, J.; Khor, H.Q.; Huang, Y.C. Shallow triple stream three-dimensional cnn (ststnet) for micro-expression recognition. In *Proceedings of the 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*; IEEE: Piscataway, NJ, USA, 2019; pp. 1–5.
8. Wei, M.; Zong, Y.; Jiang, X.; Lu, C.; Liu, J. Micro-expression recognition using uncertainty-aware magnification-robust networks. *Entropy* **2022**, *24*, 1271. [\[CrossRef\]](#) [\[PubMed\]](#)
9. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. In *Proceedings of the ICLR 2021, Virtual Event, Austria, 3–7 May 2021*.
10. Wei, M.; Zheng, W.; Jiang, X.; Zong, Y.; Lu, C.; Liu, J. A novel magnification-robust network with sparse self-attention for micro-expression recognition. In *Proceedings of the 2022 26th International Conference on Pattern Recognition (ICPR)*; IEEE: Piscataway, NJ, USA, 2022; pp. 1120–1126.
11. Fan, X.; Chen, X.; Jiang, M.; Shahid, A.R.; Yan, H. Selfme: Self-supervised motion learning for micro-expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, BC, Canada, 17–24 June 2023; pp. 13834–13843.
12. Zhai, Z.; Zhao, J.; Long, C.; Xu, W.; He, S.; Zhao, H. Feature representation learning with adaptive displacement generation and transformer fusion for micro-expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, BC, Canada, 17–24 June 2023; pp. 22086–22095.
13. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 213–229.
14. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.

15. Shen, Z.; Zhang, M.; Zhao, H.; Yi, S.; Li, H. Efficient attention: Attention with linear complexities. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Virtual, 5–9 January 2021; pp. 3531–3539.
16. Kitaev, N.; Kaiser, L.; Levskaya, A. Reformer: The Efficient Transformer. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 26 April–1 May 2020.
17. Xia, Z.; Pan, X.; Song, S.; Li, L.E.; Huang, G. Vision transformer with deformable attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 4794–4803.
18. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 764–773.
19. Zhou, L.; Mao, Q.; Dong, M. Objective Class-Based Micro-Expression Recognition Through Simultaneous Action Unit Detection and Feature Aggregation. *Tsinghua Sci. Technol.* **2025**, *30*, 2114–2132. [\[CrossRef\]](#)
20. Li, Y.; Huang, X.; Zhao, G. Micro-expression action unit detection with spatial and channel attention. *Neurocomputing* **2021**, *436*, 221–231. [\[CrossRef\]](#)
21. Varanka, T.; Peng, W.; Zhao, G. Micro-expression recognition with noisy labels. *Electron. Imaging* **2021**, *33*, 1–8. [\[CrossRef\]](#)
22. Wang, T.; Shang, L. Temporal augmented contrastive learning for micro-expression recognition. *Pattern Recognit. Lett.* **2023**, *167*, 122–131. [\[CrossRef\]](#)
23. Ruan, B.K.; Lo, L.; Shuai, H.H.; Cheng, W.H. Mimicking the annotation process for recognizing the micro expressions. In Proceedings of the 30th ACM International Conference on Multimedia, Lisbon, Portugal, 10–14 October 2022; pp. 228–236.
24. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
25. Yu, J.; Zhao, G.; Zhang, Y.; He, P.; Zhang, Z.; Yang, Z.; Liu, Q.; Sun, J.; Liang, J. 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, Melbourne, Australia, 28 October–1 November 2024; pp. 11484–11489.
26. Tong, Z.; Song, Y.; Wang, J.; Wang, L. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 10078–10093. [\[CrossRef\]](#)
27. Zaheer, M.; Guruganesh, G.; Dubey, K.A.; Ainslie, J.; Alberti, C.; Ontanon, S.; Pham, P.; Ravula, A.; Wang, Q.; Yang, L.; et al. Big bird: Transformers for longer sequences. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 17283–17297.
28. Bo, W.; Zhang, S.; Sun, Y.; Wu, J.; Xie, Q.; Tan, X.; Chen, K.; He, W.; Li, X.; Zhao, N.; et al. ViLoMem: Agentic Learner with Grow-and-Refine Multimodal Semantic Memory. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Denver, CO, USA, 3–7 June 2026; pp. 5476–5486.
29. Wu, X.; Liang, D.; Feng, T.; Xia, K.; Zhang, Y.; Li, X.; Tan, X.; Bai, X. Generation Models Know Space: Unleashing Implicit 3D Priors for Scene Understanding. In Proceedings of the European Conference on Computer Vision (ECCV), Malmö, Sweden, 8–12 September 2026.
30. Buolamwini, J.; Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the Conference on Fairness, Accountability and Transparency*; PMLR: Piscataway, NJ, USA, 2018; pp. 77–91.
31. Datta, A.; Swamidass, S.J. Fair-Net: A Network Architecture for Reducing Performance Disparity between Identifiable Subpopulations. In *Proceedings of the 14th International Conference on Agents and Artificial Intelligence (ICAART 2022)*, Vienna, Austria, 3–5 February 2022; Science and Technology Publications, Lda: Setúbal, Portugal, 2022; Volume 3, pp. 645–654.
32. Zhang, Y.; Xu, X.; Zhao, Y.; Wen, Y.; Tang, Z.; Liu, M. Facial prior guided micro-expression generation. *IEEE Trans. Image Process.* **2023**, *33*, 525–540. [\[CrossRef\]](#)
33. See, J.; Yap, M.H.; Li, J.; Hong, X.; Wang, S.J. Megc 2019—the second facial micro-expressions grand challenge. In *Proceedings of the 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*; IEEE: Piscataway, NJ, USA, 2019; pp. 1–5.
34. Yan, W.J.; Li, X.; Wang, S.J.; Zhao, G.; Liu, Y.J.; Chen, Y.H.; Fu, X. CASME II: An improved spontaneous micro-expression database and the baseline evaluation. *PLoS ONE* **2014**, *9*, e86041. [\[CrossRef\]](#) [\[PubMed\]](#)
35. Davison, A.K.; Lansley, C.; Costen, N.; Tan, K.; Yap, M.H. Samm: A spontaneous micro-facial movement dataset. *IEEE Trans. Affect. Comput.* **2016**, *9*, 116–129. [\[CrossRef\]](#)
36. Li, X.; Pfister, T.; Huang, X.; Zhao, G.; Pietikäinen, M. A spontaneous micro-expression database: Inducement, collection and baseline. In *Proceedings of the 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*; IEEE: Piscataway, NJ, USA, 2013; pp. 1–6.
37. Varanka, T.; Li, Y.; Peng, W.; Zhao, G. Data leakage and evaluation issues in micro-expression analysis. *IEEE Trans. Affect. Comput.* **2023**, *15*, 186–197. [\[CrossRef\]](#)
38. Yan, W.J.; Wu, Q.; Liu, Y.J.; Wang, S.J.; Fu, X. CASME database: A dataset of spontaneous micro-expressions collected from neutralized faces. In *Proceedings of the 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*; IEEE: Piscataway, NJ, USA, 2013; pp. 1–7.

39. Li, J.; Dong, Z.; Lu, S.; Wang, S.J.; Yan, W.J.; Ma, Y.; Liu, Y.; Huang, C.; Fu, X. CAS (ME) 3: A third generation facial spontaneous micro-expression database with depth information and high ecological validity. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 2782–2800.
40. Ben, X.; Ren, Y.; Zhang, J.; Wang, S.J.; Kpalma, K.; Meng, W.; Liu, Y.J. Video-based facial micro-expression analysis: A survey of datasets, features and algorithms. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 5826–5846.
41. Li, X.; Cheng, S.; Li, Y.; Behzad, M.; Shen, J.; Zafeiriou, S.; Pantic, M.; Zhao, G. 4DME: A spontaneous 4D micro-expression dataset with multimodalities. *IEEE Trans. Affect. Comput.* **2022**, *14*, 3031–3047. [[CrossRef](#)]
42. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022.
43. Zhao, G.; Pietikainen, M. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 915–928. [[CrossRef](#)] [[PubMed](#)]
44. Nie, X.; Takalkar, M.A.; Duan, M.; Zhang, H.; Xu, M. GEME: Dual-stream multi-task GEndEr-based micro-expression recognition. *Neurocomputing* **2021**, *427*, 13–28. [[CrossRef](#)]
45. Zhao, S.; Tao, H.; Zhang, Y.; Xu, T.; Zhang, K.; Hao, Z.; Chen, E. A two-stage 3D CNN based learning method for spontaneous micro-expression recognition. *Neurocomputing* **2021**, *448*, 276–289. [[CrossRef](#)]
46. Zhang, L.; Hong, X.; Arandjelović, O.; Zhao, G. Short and long range relation based spatio-temporal transformer for micro-expression recognition. *IEEE Trans. Affect. Comput.* **2022**, *13*, 1973–1985. [[CrossRef](#)]
47. Wang, Z.; Yang, M.; Jiao, Q.; Xu, L.; Han, B.; Li, Y.; Tan, X. Two-level spatio-temporal feature fused two-stream network for micro-expression recognition. *Sensors* **2024**, *24*, 1574. [[CrossRef](#)] [[PubMed](#)]
48. Gan, Y.S.; Lien, S.E.; Chiang, Y.C.; Liong, S.T. LAENet for micro-expression recognition. *Vis. Comput.* **2024**, *40*, 585–599.
49. Yang, J.; Wu, Z.; Wu, R. Micro-expression recognition based on contextual transformer networks. *Vis. Comput.* **2025**, *41*, 1527–1541.
50. Yu, Z.; Chen, X.; Qu, C. SDGSA: A lightweight shallow dual-group symmetric attention network for micro-expression recognition. *Complex Intell. Syst.* **2024**, *10*, 8143–8162. [[CrossRef](#)]
51. Wang, Z.; Zhang, K.; Luo, W.; Sankaranarayana, R. Htnet for micro-expression recognition. *Neurocomputing* **2024**, *602*, 128196. [[CrossRef](#)]
52. Liu, Y.J.; Zhang, J.K.; Yan, W.J.; Wang, S.J.; Zhao, G.; Fu, X. A main directional mean optical flow feature for spontaneous micro-expression recognition. *IEEE Trans. Affect. Comput.* **2015**, *7*, 299–310. [[CrossRef](#)]
53. Liu, Y.; Du, H.; Zheng, L.; Gedeon, T. A neural micro-expression recognizer. In *Proceedings of the 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*; IEEE: Piscataway, NJ, USA, 2019; pp. 1–4.
54. Xia, Z.; Peng, W.; Khor, H.Q.; Feng, X.; Zhao, G. Revealing the invisible with model and data shrinking for composite-database micro-expression recognition. *IEEE Trans. Image Process.* **2020**, *29*, 8590–8605. [[CrossRef](#)]
55. Tran, D.; Wang, H.; Torresani, L.; Ray, J.; LeCun, Y.; Paluri, M. A closer look at spatiotemporal convolutions for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6450–6459.
56. Zhou, T.; Liu, J.; Jin, Y.; Yao, L. MicroMamba: State Space Model with Partitioned Window Scan for Micro-Expression Recognition. In Proceedings of the 6th ACM International Conference on Multimedia in Asia, Auckland, New Zealand, 3–6 December 2024; pp. 1–7.
57. Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.