Uncovering Hidden Emotions with Adaptive Multi-Attention Graph Networks

Ankith Jain Rakesh Kumar and Bir Bhanu
Department of Electrical and Computer Engineering
University of California, Riverside
arake001@ucr.edu, bhanu@ece.ucr.edu

Abstract

Micro-expressions (MEs) are subtle expressions lasting a fraction of a second, offering valuable cues for understanding human emotions and intentions. However, effectively classifying these subtle expressions from video data poses several challenges due to their short duration and low intensity. This paper addresses these issues and presents a novel 2-stream Adaptive Multi-Attention ((Self-Attention and Gaussian Attention) Graph Network (2S-AMAGN) based approach for ME classification in videos. The Self-Attention mechanism captures the global and local dependencies between nodes in a graph. The Gaussian attention mechanism computes weights based on the Gaussian distribution, considering the mean and variance of features across each edge, offering a nuanced understanding of spatial and temporal relationships within MEs. It meticulously analyzes node pair features and edge features, capturing the significance of facial regions. An adaptive learnable weight is introduced to learn the contributions of each attention mechanism, facilitating adaptive attention fusion. The network utilizes a three-frame graph structure to extract spatio-temporal information. The approach incorporates a dynamic frame selection mechanism, which utilizes a sliding window optical flow method to filter out low-intensity emotion frames, thereby refining the extraction of spatio-temporal information from the video data. The results are presented and compared with state-of-the-art methods for SMIC and SAMM databases. Additionally, cross-dataset experiments are conducted, and the results are reported.

1. Introduction

Facial expressions convey information about human emotions, intentions, and social interactions. They serve as a universal language, enabling individuals to communicate feelings nonverbally. Understanding facial expressions is crucial for effective communication, social interaction, and empathy. In addition, facial expression analysis plays a vital role in various fields, including psychology, human-computer interaction, affective computing, and forensics.

Facial expressions can be classified into two main categories: macro-expression and micro-expression. Macro-expressions, lasting more than 1 second, involve voluntary movements and noticeable changes across a wide facial area, making them easily recognizable by both humans and machines. In contrast, micro-expressions are subtle, brief, and spontaneous, and the duration of these expressions is below 0.6s [1]. These subtle manifestations, often invisible to the untrained eye, carry rich information about an individual's true emotional state. By decoding the subtle cues embedded in MEs, researchers can gain insights into human behavior, improve emotional intelligence in machines, and enhance the quality of human-computer interaction.

Micro-expression (ME) classification has traditionally relied on conventional methods such as Bi-WOOF [2], LBP-TOP [3], optical flow [4], optical strain [5], and 3DHOG [6] to capture spatial and temporal information. More recently, researchers have explored the use of Convolutional Neural Networks (CNNs) [7, 8] and Graph Neural Networks (GNNs) [9, 10] for ME classification. For instance, Xie *et al.* [9] utilized AU features and self-attention graph networks, while Kumar *et al.* [10] employed landmark points and optical flow features with self-attention mechanisms using graph networks. Despite these advancements, these approaches often face challenges in extracting subtle features from video frames and accurately capturing spatio-temporal information.

The classification of MEs is challenging due to their subtle behavior and brief duration. Moreover, there is a shortage of large and balanced datasets, complicating the training of end-to-end CNN, Transformers, and GNN models.

Various attention mechanisms are employed in deep learning applications, such as spatial attention [11] and self-attention [10]. Traditionally, ME classification tasks have relied solely on self-attention. However, a key motivation for our approach, which integrates the Adaptive Multi-Attention mechanism, is the limitation of self-attention in capturing complex spatial relationships and subtle variations inherent in MEs

In order to address the above challenges, we propose a novel 2-Stream Adaptive Multi-Attention Graph Network (2S-AMAGN), incorporating Self-Attention and Gaussian Attention mechanisms, to efficiently learn node, node pair, and edge features. Self-attention mechanism facilitates the analysis of local and global dependencies between nodes, while Gaussian attention computes weights for each node pair by considering the mean and variance of node and edge features via the Gaussian distribution. We employ an adaptive learnable weight to compute the contributions of each attention mechanism and subsequently fuse them. This fusion of attention scores enhances the weighting of each node and its features, enabling effective capture of subtle changes on the human face. We adopt Self-Attention Graph Pooling (SAGPOOL), leveraging a self-attention network to compute confidence scores for each node. Finally, we fuse the two-stream graph network for the classification of MEs.

We design a three-frame graph structure to capture spatio-temporal information, incorporating node location features and optical flow patch features as node features for the 2-stream network. Additionally, we employ Jaccard's index and the radial basis function to compute edge features. We implement a sliding window optical flow approach to remove low-intensity expression frames from video. Moreover, we address data imbalance by augmenting training data samples using multiple amplification factors of the Eulerian Motion Magnification (EMM) method, particularly for expression categories with limited video data. We conduct ablation analysis to assess the significance of each component of our technique and perform cross-dataset experiments to evaluate its generalization capabilities.

2. Related Work and Contributions

2.1. Related Work

Over the past decade, researchers in computer vision and cognitive psychology have directed their attention toward spotting and classifying MEs. Our current research is centered around classifying MEs. There are various approaches involved in the extraction of features for the classification of MEs, such as (i) handcrafted approaches, (ii) CNNs and Transformers, and (iii) GNNs, as shown in Table 1.

The first class of methods comprises handcrafted approaches, as shown in Table 1 such as LBP-TOP, Bi-WOOF, HOG, optical flow, and optical strain. These approaches are unable to capture subtle variations in facial regions and also lack in computing spatio-temporal information.

The second class of approaches comprises CNNs and Transformer networks, as shown in Table 1. CNN approaches encounter challenges in capturing both local and global interactions between facial regions. Transformer networks face issues due to their dependency on large datasets and high computational resources.

The third class of techniques comprises of GNNs, as shown in Table 1. GNNs excel in capturing the subtle nuances present in MEs and are adept at capturing both local and global interactions among facial regions. Therefore, we employ GNNs in our approach to classify MEs.

2.2. Contributions

The contributions of this paper are:

- We present a landmark-assisted 2-stream Adaptive Multi-Attention Graph Network, which uses a Self-Attention and Gaussian attention mechanism to learn the dependencies within individual nodes and the relationships between pairs of nodes.
- We design an adaptive learnable weight to compute the contributions of each attention mechanism and adaptively fuse the attention scores.
- We employ an adaptive frame selection approach, utilizing a sliding window optical flow method, to identify and retain frames exhibiting high intensity of expression while discarding those with low intensity from the video.
- We conduct a comprehensive evaluation of our approach on two available datasets (SMIC and SAMM), covering MEs across three and five categories. Additionally, we assess the performance of our method on cross-datasets to evaluate its generalization capabilities.

3. Technical Approach

The methodology employed for classifying MEs is shown in Fig. 1. To enhance the input videos, Eulerian motion magnification (EMM) [25] is utilized. Following this, a sliding window optical flow approach is applied to segment the videos, enabling the removal of low-intensity expression frames while utilizing the remaining high-intensity frames. Landmark points are computed using the dlib [26] software and, together with optical flow patch features, serve as node features. These node features are complemented by local and global edge features to construct our graph. Our approach incorporates a two-stream adaptive learned multi-attention network with Self-Attention and Gaussian Attention mechanisms, facilitating the categorization of MEs and leveraging Self-Attention Graph Pooling (SAGPOOL) [27]. Finally, features from both streams of the network are concatenated and passed through a fully connected and softmax layer for ME classification.

3.1. Frame Selection Approach

We partition the video frames into multiple segments using a sliding window approach with each segment containing eight frames. Subsequently, we compute the optical flow between each frame and its consecutive frame within the segment. The optical flow components for each frame are then summed, and the average optical flow for each segment

Table 1. Research studies focusing on the use of features for classifying MEs

Author	Year	Technique	Video/Image Frames	Attributes Extractor	Classifier
Saeed *et al.* [12]	2021	Handcrafted	Video	LGBP + LBP-TOP	SVM
Liong *et al.* [6]	2018	Handcrafted	Video	Optical strain	SVM
Gan *et al.* [7]	2019	CNN	Onset + Apex	Optical Flow + CNN	MLP
Khor *et al.* [4]	2018	CNN	Video	Optical flow + CNN-LSTM	SVM
Kumar *et al.* [13]	2021	CNN	Video	CNN, CNN-LSTM, 3DHOG	SVM, MLP
Khor *et al.* [14]	2019	CNN	Video	2S-CNN	MLP
Song *et al.* [15]	2019	CNN	Onset, Apex and Offset	3S-CNN	MLP
Wang *et al.* [16]	2023	CNN	Video	Contrastive Learning	MLP
Guo *et al.* [17]	2023	Transformer	Video	3DCNN + Transformer	MLP
Fan *et al.* [18]	2023	Transformer	Onset and Apex	Contrastive Vision Transformer	MLP
Wang *et al.* [19]	2024	Transformer	Onset and Apex	Modified ResNet and Transformer	MLP
Lo *et al.* [20]	2020	GNN	Video	AU + 3D CNN + GNN	MLP
Xie *et al.* [9]	2020	GNN	Video	AU + GCN	MLP
Zhou *et al.* [21]	2020	GNN	Onset + Apex	Optical flow + AU + GCN	MLP
Kumar *et al.* [10]	2021	GNN	Frames	Landmark points + Optical flow + GAT	MLP
Lei *et al.* [22]	2021	GNN	Video	CNN + Graph transformer	MLP
Kumar *et al.* [23]	2022	GNN	High-Intensity Frames	Landmark points + Optical flow + GAT	MLP
Kumar *et al.* [24]	2023	GNN	High-Intensity Frames	Node Features + Edge features + EdgeNode GAT	MLP

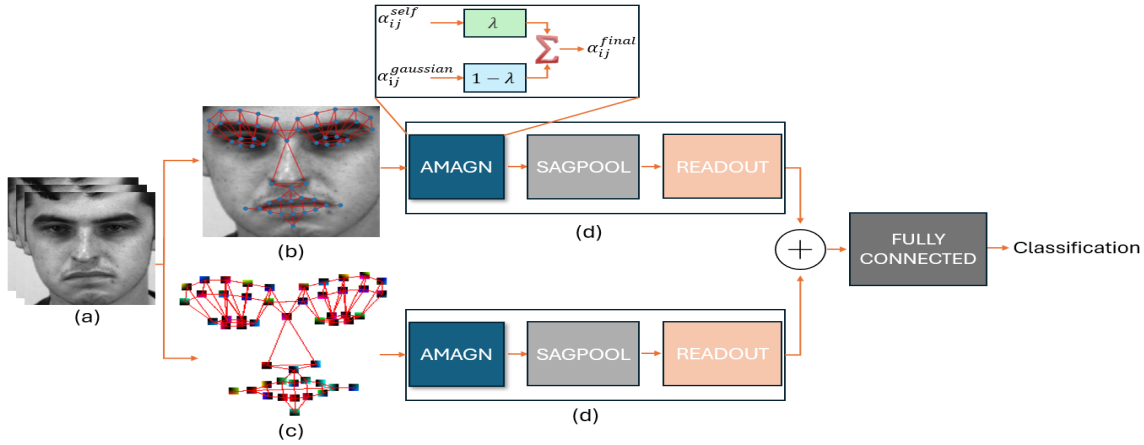


Figure 1. Overview of our proposed 2-stream Adaptive Multi-Attention Graph Network (2S-AMAGN) approach: (a) Magnified input video using EMM [25]; (b) Node location features with edge features; (c) Optical flow patch node features with edge features; (d) Adaptive Multi-Attention Graph Network (AMAGN) with three layers, enlarged to show its workings within a box alongside the Self-Attention Graph Pooling (SAGPOOL) layer and Readout layer. The output of both streams is concatenated and passed through the fully connected layer and softmax layer for the classification of MEs. Here, '+' indicates the concatenation operation, while Σ represents the summation of α_{ij}^{self} (Self-Attention mechanism) and $\alpha_{ij}^{gaussian}$ (Gaussian attention mechanism) learned with the confidence score λ. α_{ij}^{final} denotes the final attention score for each node pair.

is computed, serving as a threshold value. If the sum of optical flow components for a frame exceeds the threshold value for its respective segment, the frame is classified as a high-intensity expression frame; otherwise, it is discarded. This process is repeated for each segment and frames classified as high-intensity expressions, and the first and last frames of the video are considered for ME classification.

3.2. Facial Graph Construction

We employ the dlib [26] software to extract the 68 landmark points on the face, from which we select only 37 points (including those representing the eyes, eyebrows, outer mouth, and some points on the nose). Additionally, we incorporate an additional 14 points, including 10 on the forehead and four near the mouth region. Consequently, each frame is represented by a total of 51 landmark points, as shown in Fig. 1 (b).

3.2.1 Selection of Node and Edge Features

In the first stream of our graph network, we utilize feature embeddings derived from location coordinate points as node features. The node feature vector size is 2, rep-

resenting the x and y coordinate positions for each node. Conversely, in the second stream, we compute optical flow information by analyzing a patch size of 10x10 surrounding the landmark coordinates. Here, the node feature vector size for the second stream is 100, as shown in Fig. 1 (c).

For edge features, we adopt the approach outlined in [24]. We employ Jaccard's similarity index to calculate the global graph structural feature, while the Radial Basis Function is utilized to compute the local graph structural feature for each edge. These edge features along with the node features play a crucial role in comprehending the intricate relationships between facial regions and their impact on the classification of MEs.

3.3. Adaptive Multi-Attention Graph Network

The Self-Attention mechanism captures global and local dependencies within individual nodes but lacks the ability to capture complex spatial relationships and subtle variations in MEs. This limitation arises from its focus on dependencies within node neighbors rather than long-range interactions between facial regions. Furthermore, the presence of noise or irrelevant information in videos can impact the performance of self-attention mechanisms, further hindering their ability to capture nuanced dynamics across different facial areas during the classification of MEs.

The Gaussian attention offers a promising solution to address the limitations of self-attention. By explicitly modeling the mean and variance of features across node pairs, Gaussian attention can better capture the spatial relationships and variations present. This allows the model to attend to informative facial regions while accounting for variations in expression intensity and spatial distribution. Therefore, integrating Gaussian attention alongside self-attention can enhance the model's capability to capture the complex spatial relationships and subtle variations inherent in MEs.

Consider a graph *G = (V, E)*, where *V* represents the vertices or nodes, and *E* represents the edges connecting the nodes. The graph consists of *N* nodes, each with node features denoted by $\mathbf{x} = \{\vec{x}_1, \vec{x}_2, \vec{x}_3, ..., \vec{x}_N\}$, where $\vec{x}_i \in R^D$, and *D* represents the total number of features in each node. Edge features are denoted by $\vec{x}_{ij}$. Following this, a graph convolutional layer computes a new set of node features as its output, denoted as $\mathbf{x}' = \{\vec{x}'_1, \vec{x}'_2, \vec{x}'_3, ..., \vec{x}'_N\}$.

3.3.1 Self-Attention Mechanism

The self-attention mechanism [28] is used in this approach. The graph convolutional layer begins by applying a learnable linear transformation using a parameterized weight matrix **W** to both node and edge features, resulting in a high-level transformation of these features. Subsequently, a self-attention mechanism is employed on the nodes and edges utilizing a shared attentional mechanism denoted as (*h*), calculated using equation (1).

$$f_{ij} = h\left(\mathbf{W}\vec{x}_i, \mathbf{W}\vec{x}_j, \mathbf{W}\vec{x}_{ij}\right) \tag{1}$$

This computation delineates the relevance of the features associated with node V_j to V_i, as well as the significance of edge features $\vec{x}_{ij}$. These f_{ij} coefficients are exclusively calculated for nodes with adjacent neighbors and edges. To ensure uniformity of coefficients across neighboring nodes and edges, the softmax function is employed for normalization, as shown in equation (2).

$$\alpha_{ij}^{self} = \frac{\exp\left(\text{LeakyReLU}\left(\vec{\mathbf{h}}^T[\mathbf{W}\vec{x}_i \| \mathbf{W}\vec{x}_j \| \mathbf{W}\vec{x}_{ij}]\right)\right)}{\sum_{k\in\mathcal{N}_i}\exp\left(\text{LeakyReLU}\left(\vec{\mathbf{h}}^T[\mathbf{W}\vec{x}_i \| \mathbf{W}\vec{x}_k \| \mathbf{W}\vec{x}_{ik}]\right)\right)} \tag{2}$$

where $\mathcal{N}_i$ represents the neighborhood of node V_i, T represents the transpose, and $||$ is the concatenation operator.

The attention mechanism h comprises a single-layer feed-forward network, parameterized by a weight vector $\vec{h}$. Following the computation of normalized attention coefficients (α_{ij}^{self}), we employ the LeakyReLU non-linear activation function. This activation function enhances the coefficients generated by the attention mechanism.

3.3.2 Gaussian Attention Mechanism

Gaussian attention is pivotal as it captures nuanced spatial-temporal relationships by modeling mean and variance, offering a more comprehensive understanding of features and enhancing the model's capability.

Gaussian attention utilizes the Gaussian probability density function to compute the mean and variance across node pairs. It involves concatenating the node features $\vec{x}_i$ and $\vec{x}_j$, along with the edge features $\vec{x}_{ij}$, and applying a linear transformation using the weight matrix $\mathbf{W}_g$. The resulting concatenated features are then averaged to obtain the mean representation μ_{ij} for the node pair *(i,j)*, as calculated using equation (3).

$$\mu_{ij} = \frac{1}{D}\sum_{k=1}^{D}\left(\mathbf{W}_g(\vec{x}_i||\vec{x}_j||\vec{x}_{ij})\right)_k \tag{3}$$

where *k* represents the index of the feature. *D* represents the total number of features.

The variance for the node pair features σ_{ij}^2 is computed using equation (4), representing the spread or dispersion of the features around the mean μ_{ij} for the node pair *(i,j)*.

$$\sigma_{ij}^2 = \frac{1}{D}\sum_{k=1}^{D}\left((\mathbf{W}_g(\vec{x}_i||\vec{x}_j||\vec{x}_{ij}) - \mu_{ij})^2\right)_k \tag{4}$$

The Gaussian attention is calculated as the attention weights $\alpha_{ij}^{gaussian}$ based on the Gaussian distribution using equation (5). Subsequently, a sum aggregation is applied to each node pair $\alpha_{ij}^{gaussian}$. Finally, the LeakyReLU and softmax function are applied using equation (6).

$$\alpha_{ij}^{gaussian} = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} \exp\left(-\frac{(\mathbf{W}_g(\vec{x}_i||\vec{x}_j||\vec{x}_{ij}) - \mu_{ij})^2}{2\sigma_{ij}^2}\right) \tag{5}$$

$$\alpha_{ij}^{gaussian} = \text{softmax}_j(\text{LeakyReLU}(\alpha_{ij}^{gaussian})) \tag{6}$$

3.3.3 Adaptive Learnable Attention Mechanism

The network adapts its use of Self-Attention and Gaussian attention mechanisms through learning, guided by the learnable parameter λ. This parameter λ (learnable parameter) regulates the balance between self-attention and Gaussian attention, shaping the final attention weight α_{ij}^{final} for each node pair, as computed using equation (7).

$$\alpha^{final} = \lambda * \alpha_{ij}^{self} + (1 - \lambda) * \alpha_{ij}^{gaussian} \tag{7}$$

3.3.4 Node Update

The final features for each node are computed using a graph convolutional operator employed to embed node features and edge features from the neighborhood. Subsequently, a non-linear activation function is applied to these embeddings. They are then aggregated to fulfill the node localization property, as shown in equation (8).

$$\vec{x}_i' = \sigma\left(\sum_{j \in \mathcal{N}_i} \alpha_{ij}^{final} \mathbf{W} \vec{x}_j\right) \tag{8}$$

where σ represents an activation function. $\vec{x}_i'$ represents the final output feature for every node.

3.4. 2-Stream Adaptive Multi-Attention Graph Network (2S-AMAGN)

We designed a novel method 2S-AMAGN aimed at extracting spatio-temporal features from videos, shown in Fig. 1. The approach involves extracting features: node locations, optical flow patches, and local and global edge features from video frames. These features are interconnected to construct a unified graph using a three-frame structure.

In the first stream of our graph network, we utilize the x and y location coordinates of landmark points as the node feature vector. This approach effectively captures the movement changes of each landmark point relative to its previous position. In the second stream, we adopt a fixed patch size for the optical flow features. The optical flow patch features component further captures spatio-temporal information about motion estimations, complementing the three-frame graph structure utilized in our network. Additionally, edge features are computed using learned node features from their respective streams within the graph network, providing supplementary insights into the relationships between nodes. The output from the final AMAGN graph layer is forwarded to the SAGPOOL layer [27]. The SAGPOOL layer effectively filters out less significant nodes based on their attention scores, employing a *top-k* selection process with a predefined ratio.

The output from the SAGPOOL layer is directed to the readout layer. Following the traversal through the readout layer in both the streams of the graph networks, the outcomes are combined by concatenation, resulting in the graph representation of the 2-stream. This combined output is subsequently passed to fully connected layers and a softmax layer for classification.

4. Experimental Results

We perform experiments on a system equipped with 128GB RAM, Intel i7 5th gen CPU, and four Titan X GPUs, and running on the Ubuntu OS 20.04.

4.1. Datasets and Preprocessing Protocols

The *two* datasets utilized for ME classification are SMIC [29] and SAMM [30]. Our objective is to categorize MEs into 3 and 5 classes. The evaluation was conducted using the Leave-One-Subject-Out Cross-Validation (LOSO-CV) approach. The video distributions for the 3 classes in the SMIC and SAMM datasets are as follows: Negative (70 and 92 videos), Positive (51 and 26 videos), and Surprise (43 and 15 videos), respectively. Similarly, for the SAMM dataset with 5 classes, the video distributions are as follows: Anger (57 videos), Happy (26 videos), Surprise (15 videos), Contempt (12 videos), and Other (26 videos). The SAMM dataset was collected at a frame rate of 200 fps, while the SMIC dataset was collected at a frame rate of 100 fps. The average number of frames per video in the SAMM dataset is 75 frames, while in the SMIC dataset, it is 34 frames.

Each image was aligned with the first frame of its corresponding video and resized to 254x254 dimensions. Our adaptive frame selection approach employs a sliding window method with optical flow analysis and a window size of 8 frames. During the training to balance the dataset classes, we varied the magnitudes of motion amplification factors between 2 and 5 to augment the video samples. For testing purposes, a consistent magnification factor of 4 was employed. In the 2S-AMAGN layer, the parameter λ is a trainable parameter whose value is determined experimentally based on the confidence score of each attention mechanism, ranging between 0 and 1. In the SAGPOOL layer, 75% of the nodes in the graph structure were retained using a ratio of k = 0.75. This retention strategy aimed to preserve essential nodes while ensuring an adequate number of nodes remained in the graph. We employed 3 Adaptive Multi-Attention layers, each with 32 hidden channels. A dropout

Table 2. Comparative performance analysis among current techniques for SAMM and SMIC datasets across three emotion classes: Positive, Negative, and Surprise. The best outcomes are highlighted in Bold, while the second-best results are denoted in Blue.

Approaches	Feature Extraction	SAMM		SMIC	
		Accuracy	UF1	Accuracy	UF1
Huang *et al.* [2016] [31]	STL-CLQP	0.6380	0.6110	-	-
Wang *et al.* [2017] [32]	LBP-TOP	0.4150	0.4060	-	-
Liong *et al.* [2018] [6]	Optical flow+BiWOOF	0.5833	0.5211	0.6159	0.5727
Khor *et al.* [2019] [14]	DSSN	0.5740	0.4640	0.6341	0.6462
Gan *et al.* [2019] [7]	CNN	0.6818	0.5423	0.6817	0.6709
Zhou *et al.* [2019] [33]	Dual Inception CNN	0.7519	0.5868	0.6585	0.6645
Kumar *et al.* [2019] [34]	CNN	0.8195	0.7056	0.7744	0.7451
Liong *et al.* [2019] [35]	3S-3DCNN	0.7744	0.6588	0.6829	0.6801
Liu *et al.* [2019] [36]	CNN	-	0.7754	-	0.7461
Xia *et al.* [2020] [37]	CNN+RCNN	0.7860	0.7410	0.7230	0.6950
Xia *et al.* [2020] [38]	CNN	-	0.6770	-	0.5980
Lo *et al.* [2020] [20]	GCN	0.5340	0.2830	-	-
Xie *et al.* [2020] [9]	AU+GAT	0.5230	0.3570	-	-
Kumar *et al.* [2021] [10]	Dual Stream GAT	0.8872	0.8118	0.7622	0.7606
Lei *et al.* [2021] [22]	Graph+AU	-	0.7751	-	0.7192
Kumar *et al.* [2022] [23]	3St-GAT	0.9098	0.8463	-	-
Zhou *et al.* [2022] [39]	FeatRef+Feature learning	0.6838	0.5436	0.7561	0.7492
Kumar *et al.* [2023] [24]	Edge-Node-GAT	-	-	0.8171	0.8143
Nguyen *et al.* [2023] [40]	Micron-BERT	-	-	-	**0.8550**
Fan *et al.* [2023] [18]	Transformer	-	-	-	0.6972
Zhai *et al.* [2023] [41]	GAN+Transformer	-	0.7720	-	0.7430
Wang *et al.* [2023] [16]	Contrastive Learning	0.6838	0.5436	0.7561	0.7492
Verma *et al.* [2023] [42]	RNAS-MER	-	0.7880	-	0.7443
Xie *et al.* [2024] [43]	CapsuleNet	-	0.7790	-	0.7848
Zhang *et al.* [2024] [44]	Transformer	0.7440	0.7370	0.7123	0.7270
Wang *et al.* [2024] [19]	CNN+Transformer	-	0.7090	-	0.7410
Ours	2S-AMAGN	**0.9323**	**0.9091**	**0.8476**	0.8508

of 0.5 was applied after the SAGPOOL layer in each stream to prevent overfitting. The optimizer used was Adam with a learning rate set to 0.001. The cross-entropy loss function was utilized in our approach.

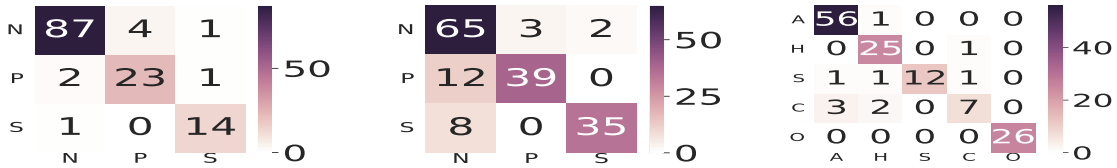
4.2. Evaluation Metrics

The distribution of data across the two datasets exposes diversity in the number of videos for various classes, with some classes being less prevalent as compared to others. As a result, relying only on the accuracy as a metric is not sufficient. To accurately assess the performance of our technique, we employ the Unweighted F1 (UF1) score alongside accuracy, as utilized in [10]. The UF1 score assigns equal significance to both infrequent and commonly occurring expression classes, making it a suitable choice for evaluation. Therefore, we opt for UF1 score and accuracy as the metrics commonly utilized in ME classification tasks.

4.3. Detailed Results

The results of our proposed 2S-AMAGN method and state-of-the-art techniques for three and five expression categories on SAMM and SMIC datasets are showcased in Table 2 and 3. Our approach exhibits enhanced performance in both accuracy and UF1-score (except slightly lower UF1 score for 3 classes for SMIC), owing to its adept feature extraction from video frames and the utilization of a 2S-AMAGN network. This enables effective discrimination between various classes of MEs.

- *SAMM Dataset (3 Classes):* Table 2 presents the results for the SAMM dataset, demonstrating that our approach 2S-AMAGN surpasses all state-of-the-art methods. For the three categories, it achieves an accuracy of 93.23% and a UF1 score of 90.91% respectively, marking an improvement of 2.25% in accuracy and 6.28% in UF1 score compared to the previous best state-of-the-art technique [23]. The confusion matrix for the SAMM database, illustrating the classification results for three categories of MEs, is shown in Fig. 2 (a).
- *SMIC Dataset (3 Classes):* The data in Table 2 showcases the results achieved with the SMIC dataset, affirming the better performance of our 2S-AMAGN method over all competing techniques concerning accuracy. Across the three categories, our approach achieves an accuracy of 84.76% and a UF1 score of 85.08%, signifying a notable advancement of 3.05% in accuracy compared to the previously leading state-of-the-art method [24]. While our



(a) SAMM datasets (3 classes) (b) SMIC datasets (3 classes) (c) SAMM datasets (5 classes)

Figure 2. Confusion matrices depict the evaluation results of classifying MEs for 3 and 5 classes across two datasets. Here, N: Negative, P: Positive, S: Surprise, A: Anger, H: Happy, C: Contempt, and O: Other category of MEs

UF1 score remains competitive, it narrowly trails behind the approach [40], which registers a UF1 score of 85.50%. The confusion matrix for the SMIC database, showcasing the classification outcomes for three categories of MEs, can be observed in Fig. 2 (b).

- *SAMM Dataset (5 Classes):* In Table 3, the results for the SAMM dataset encompassing five classes of MEs reveal the better performance of our approach, 2S-AMAGN, over all the state-of-the-art methods. Across the five categories, it achieves an accuracy of 92.65% and a UF1 score of 88.44%, showcasing a notable enhancement of 2.94% in accuracy compared to the previously best state-of-the-art technique [23]. Similarly, our approach demonstrates an improvement of 4.58% in terms of UF1 score compared to the previous best method [40]. The confusion matrix for the SAMM database, delineating the classification outcomes for five categories of MEs, is illustrated in Fig. 2 (c).

4.4. Ablation Study Results

We examined the efficacy of our proposed method through an ablation study, evaluating the influence of self-attention versus the multi-attention network. We interpreted the performance of the attention mechanisms, and the findings are detailed in Tables 4 and 5, illustrating results for three and five categories of MEs, respectively.

Tables 4 and 5 demonstrate notable enhancements in accuracy and UF1 score for both SAMM and SMIC datasets with the multi-attention approach. On the SAMM dataset 3 classes, there is a 2.25% improvement in accuracy and a remarkable 6.93% improvement in UF1 score compared to the self-attention approach. Similarly, for the SMIC dataset, the multi-attention mechanism achieves a 3.05% accuracy improvement and a substantial 4.22% increase in UF1 score over the self-attention approach. Likewise, for the SAMM dataset 5 classes, there is a 2.21% improvement in accuracy and a 3.98% improvement in UF1 score compared to the self-attention approach.

The proposed adaptive multi-attention mechanism dynamically integrates self-attention and Gaussian attention to effectively process MEs. This adaptive approach adjusts its focus based on the contextual structure of the input graph, ensuring the most relevant information is captured.

Table 3. Comparative performance analysis of current techniques for SAMM datasets for 5 classes of emotions. The best results are in Bold and the second-best results are in Blue

Approaches	Feature Extractor	Accuracy	UF1
Khor *et al.* [2019] [14]	CNN	0.5294	0.4260
Khor *et al.* [2019] [14]	SSSN	0.5662	0.4513
Khor *et al.* [2019] [14]	DSSN	0.5735	0.4644
Song *et al.* [2019] [15]	3S-CNN	0.7176	0.6942
Xia *et al.* [2020] [45]	2S-CNN+GAN	0.7410	0.7360
Li *et al.* [2021] [46]	CNN+Att.	0.4090	0.3400
Su *et al.* [2021] [47]	2S-CNN+Att.	0.6324	0.5709
Nie *et al.* [2021] [48]	2S-CNN+ML	0.5588	0.4538
Kumar *et al.* [2021] [10]	GACNN	0.8824	0.8279
Lei. *et al.* [2021] [49]	Graph TCN	0.7500	0.6985
Lei *et al.* [2021] [22]	Graph-AU	0.7426	0.7045
Kumar *et al.* [2022] [23]	3St+GAT	0.8971	0.8365
Nguyen *et al.* [2023] [40]	Micron-BERT	-	0.8386
Feng *et al.* [2023] [50]	KPCANet	0.6383	0.5215
Ours	2S-AMAGN	**0.9265**	**0.8844**

Gaussian attention assesses the mean and variance of node and edge features across each edge, allowing the model to capture subtle nuances and spatial distribution within facial regions with greater precision. This statistical focus on the nuances of node and edge features leads to a more targeted analysis of subtle emotional changes. Meanwhile, self-attention captures essential node and edge features by recognizing local and global dependencies, revealing subtle connections and temporal patterns that might otherwise be overlooked. By evaluating the significance of each node and the relationships between them, the model can uncover key patterns and interactions that enhance the understanding of subtle emotional cues in MEs. The interaction between these two attention mechanisms enhances the model's sensitivity to minute variations and dynamic shifts within MEs, allowing for a more comprehensive understanding of complex emotional cues. This balanced integration adaptively combines insights from both spatial and temporal relationships, leading to improved classification accuracy and UF1 score. By leveraging the full spectrum of information encoded in the data, the approach offers a nuanced and robust

method for analyzing MEs.

Table 4. Ablation study results for SAMM and SMIC databases for 3 classes of emotions.

Network	SAMM		SMIC	
	Accuracy	UF1	Accuracy	UF1
Self-Attention	0.9098	0.8398	0.8171	0.8086
Multi-Attention	**0.9323**	**0.9091**	**0.8476**	**0.8508**

Table 5. Ablation study results for SAMM database for 5 classes of emotions.

Network	SAMM	
	Accuracy	UF1
Self-Attention	0.9044	0.8446
Multi-Attention	**0.9265**	**0.8844**

4.5. Cross-Dataset Evaluation Results

To demonstrate the effectiveness of our approach across different scenarios and with participants of diverse genders, races, and ages, we conducted cross-dataset evaluations. We utilized the methodology outlined in Section 3 for this analysis. The outcomes of the cross-dataset evaluation, focusing on three classes of MEs, are presented in Table 6.

The outcomes of the cross-dataset evaluation on three classes of MEs are presented in Table 6. When trained on the SAMM dataset and tested on the SMIC dataset, we achieved an accuracy of *76.22*% and a UF1-Score of *73.92*%. Conversely, training on the SMIC dataset and evaluating on the SAMM dataset yielded an accuracy of *88.72*% and a UF1 Score of *79.27*%.

The results of the cross-dataset evaluation highlight the adaptability of our model across diverse datasets and its ability to generalize effectively to new data. Our proposed method demonstrates consistent performance across various environments and with participants from diverse backgrounds, irrespective of age, gender, or ethnicity. Specifically, the approach exhibits proficiency in accurately classifying MEs into three distinct emotion classes.

4.6. Computational Time Analysis

The model consists of 0.038 million parameters and utilizes around 4GB of GPU memory. Our approach is divided into three key steps: (i) reading video frames and removing low-intensity expression frames from a video using a sliding window optical flow approach with a window size of 8 frames. This process takes $\sim$ 1.05s for a 200 fps video and $\sim$ 0.52s for a 100 fps video. The output of this step is a selection of high-intensity expression frames. (ii) The face graph construction step involves detecting the landmark points as node location features in each frame, computing optical flow features for each patch around each node, and calculating edge features. Finally, constructing the entire graph structure for the video takes $\sim$ 0.4s for a 200 fps video and $\sim$ 0.2s for a 100 fps video. The output of this step is the graph structure of the video in Pytorch Geometric format. (iii) Inference/testing time is $\sim$ 0.03s per video, resulting in an overall time to classify a 100 fps video of $\sim$ 0.75s, while a 200 fps video takes $\sim$ 1.48s.

Table 6. Cross dataset examination on two micro-expression databases (3 classes of emotions).

Training Database	Evaluating Database			
	SAMM		SMIC	
	Accuracy	UF1	Accuracy	UF1
SAMM	-	-	0.7622	0.7392
SMIC	0.8872	0.7927	-	-

5. Conclusions and Future Work

In this paper, we introduced a novel approach called 2-Stream Adaptive Multi-Attention Graph Network (2S-AMAGN). This approach integrated adaptive learnable attention on graph data from both self-attention and Gaussian attention mechanisms, utilizing node location and optical flow patch information as node features, and local and global edge features. The network efficiently learned the distribution of features and variations within each ME video through the Gaussian attention mechanism, while also capturing local and global interactions between nodes using the self-attention mechanism. By learning the contributions of each attention mechanism, the network dynamically fused them to effectively classify MEs. Additionally, we introduced a dynamic frame selection using a sliding window approach based on optical flow information. We constructed a 3 frame graph structure to capture the spatio-temporal information. We conducted extensive evaluations involving two ME datasets, an ablation study analysis, and cross-dataset evaluations. Our approach demonstrated an average improvement of 2.65% and 2.93% in UF1 score across 3 ME classes (SAMM and SMIC). For SAMM's 5 ME classes, we achieved a notable improvement of 2.94% in accuracy and 4.58% in UF1 score. The cross-dataset evaluation results confirmed the effectiveness of our approach in various scenarios and with diverse subjects, irrespective of age, gender, or ethnicity differences. In the future, we aim to explore insights into multi-attention for spatial and temporal relationships these mechanisms are most effective at modeling.

6. Acknowledgments

This material is based upon work supported by the National Science Foundation under grant number 1911197.

References

[1] M. Shreve, S. Godavarthy, D. Goldgof, and S. Sarkar, "Macro-and micro-expression spotting in long videos using spatio-temporal strain," in *IEEE Face and Gesture*, 2011. 1

[2] A. K. Davison, W. Merghani, and M. H. Yap, "Objective classes for micro-facial expression recognition," *Journal of Imaging*, vol. 4, 2018. 1

[3] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, 2007. 1

[4] H. Khor, J. See, R. C. W. Phan, and W. Lin, "Enriched long-term recurrent convolutional network for facial micro-expression recognition," in *13th IEEE International Conference on Automatic Face Gesture Recognition*, May 2018. 1, 3

[5] S. T. Liong, J. See, R. C.-W. Phan, Y.-H. Oh, A. Cat Le Ngo, K. Wong, and S.-W. Tan, "Spontaneous subtle expression detection and recognition based on facial strain," *Signal Processing: Image Communication*, vol. 47, 2016. 1

[6] S. Liong, J. See, K. Wong, and R. C. W. Phan, "Less is more: Micro-expression recognition from video using apex frame," *Signal Processing: Image Communication*, vol. 62, 2018. 1, 3, 6

[7] Y. Gan, S. T. Liong, W. C. Yau, Y. C. Huang, and L. K. Tan, "OFF-ApexNet on micro-expression recognition system," *Signal Processing: Image Communication*, vol. 74, 2019. 1, 3, 6

[8] M. Peng, C. Wang, T. Chen, G. Liu, and X. Fu, "Dual temporal scale convolutional neural network for micro-expression recognition," *Frontiers in Psychology*, vol. 8, 2017. 1

[9] H. X. Xie, L. Lo, H. H. Shuai, and W. H. Cheng, "AU-assisted graph attention convolutional network for micro-expression recognition," in *28th ACM International Conference on Multimedia*, 2020. 1, 3, 6

[10] A. J. R. Kumar and B. Bhanu, "Micro-expression classification based on landmark relations with graph attention convolutional network," in *Proceedings of the IEEE/CVF Conference on CVPR Workshops*, 2021. 1, 3, 6, 7

[11] F. F. Niloy, M. A. Amin, A. A. Ali, and A. M. Rahman, "Attention toward neighbors: A context aware framework for high resolution image segmentation," in *2021 IEEE International Conference on Image Processing (ICIP)*, 2021, pp. 2279–2283. 1

[12] U. Saeed, "Facial micro-expressions as a soft biometric for person recognition," *Pattern Recognition Letters*, vol. 143, 2021. 3

[13] A. J. Rakesh Kumar, B. Bhanu, C. Casey, S. G. Cheung, and A. Seitz, "Depth videos for the classification of micro-expressions," in *25th International Conference on Pattern Recognition (ICPR)*, 2021. 3

[14] H. Khor, J. See, S. Liong, R. C. W. Phan, and W. Lin, "Dual-stream shallow networks for facial micro-expression recognition," in *IEEE International Conference on Image Processing (ICIP)*, 2019. 3, 6, 7

[15] B. Song, K. Li, Y. Zong, J. Zhu, W. Zheng, J. Shi, and L. Zhao, "Recognizing spontaneous micro-expression using a three-stream convolutional neural network," *IEEE Access*, vol. 7, 2019. 3, 7

[16] T. Wang and L. Shang, "Temporal augmented contrastive learning for micro-expression recognition," *Pattern Recognition Letters*, vol. 167, 2023. 3, 6

[17] X. Guo, X. Zhang, L. Li, and Z. Xia, "Micro-expression spotting with multi-scale local transformer in long videos," *Pattern Recognition Letters*, 2023. 3

[18] X. Fan, X. Chen, M. Jiang, A. R. Shahid, and H. Yan, "Selfme: Self-supervised motion learning for micro-expression recognition," in *Proceedings of the IEEE/CVF CVPR*, June 2023. 3, 6

[19] Z. Wang, M. Yang, Q. Jiao, L. Xu, B. Han, Y. Li, and X. Tan, "Two-level spatio-temporal feature fused two-stream network for micro-expression recognition," *Sensors*, vol. 24, no. 5, 2024. 3, 6

[20] L. Lo, H. Xie, H. Shuai, and W. Cheng, "MER-GCN: Micro-expression recognition based on relation modeling with graph convolutional networks," in *IEEE Conference on Multimedia Information Processing and Retrieval (MIPR),*, 2020. 3, 6

[21] L. Zhou, Q. rong Mao, and M. Dong, "Objective class-based micro-expression recognition through simultaneous action unit detection and feature aggregation," *ArXiv*, vol. abs/2012.13148, 2020. 3

[22] L. Lei, T. Chen, S. Li, and J. Li, "Micro-expression recognition based on facial graph representation learning and facial action unit fusion," in *Proceedings of the IEEE/CVF Conference on CVPR Workshops*, 2021. 3, 6, 7

[23] A. J. Rakesh Kumar and B. Bhanu, "Three stream graph attention network using dynamic patch selection for the classification of micro-expressions," in *2022 IEEE/CVF Conference on CVPR Workshops*, 2022. 3, 6, 7

[24] A. J. R. Kumar and B. Bhanu, "Relational edge-node graph attention network for classification of micro-expressions," in *Proceedings of the IEEE/CVF CVPR Workshops*, 2023. 3, 4, 6

[25] H. Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. T. Freeman, "Eulerian video magnification for revealing subtle changes in the world," *ACM Transactions on Graphics (SIGGRAPH)*, vol. 31, 2012. 2, 3

[26] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, 2009. 2, 3

[27] J. Lee, I. Lee, and J. Kang, "Self-attention graph pooling," in *Proceedings of the 36th International Conference on Machine Learning*, vol. 97. PMLR, 2019. 2, 5

[28] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *International Conference on Learning Representations*, 2018. 4

[29] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikäinen, "A spontaneous micro-expression database: Inducement, collection and baseline," in *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 2013. 5

[30] A. K. Davison, C. Lansley, N. Costen, K. Tan, and M. H. Yap, "SAMM: A spontaneous micro-facial movement dataset," *IEEE Transactions on Affective Computing*, vol. 9, 2018. 5

[31] X. Huang, G. Zhao, X. Hong, W. Zheng, and M. Pietikäinen, "Spontaneous facial micro-expression analysis using spatiotemporal completed local quantized patterns," *Neurocomputing*, vol. 175, 2016. 6

[32] Y. Wang, J. See, Y.-H. Oh, R. C.-W. Phan, Y. Rahulamathavan, H.-C. Ling, S.-W. Tan, and X. Li, "Effective recognition of facial micro-expressions with video motion magnification," *Multimedia Tools Appl.*, vol. 76, Oct. 2017. 6

[33] L. Zhou, Q. Mao, and L. Xue, "Dual-inception network for cross-database micro-expression recognition," in *14th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2019. 6

[34] A. J. R. Kumar, R. Theagarajan, O. Peraza, and B. Bhanu, "Classification of facial micro-expressions using motion magnified emotion avatar images," in *IEEE Conference on CVPR Workshops*, 2019. 6

[35] S. T. Liong, Y. S. Gan, J. See, H. Q. Khor, and Y. C. Huang, "Shallow triple stream three-dimensional cnn (ststnet) for micro-expression recognition," in *14th IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, 2019. 6

[36] Y. Liu, H. Du, L. Zheng, and T. Gedeon, "A neural micro-expression recognizer," in *14th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2019. 6

[37] Z. Xia, X. Hong, X. Gao, X. Feng, and G. Zhao, "Spatiotemporal recurrent convolutional networks for recognizing spontaneous micro-expressions," *IEEE Transactions on Multimedia*, vol. 22, no. 3, 2020. 6

[38] Z. Xia, W. Peng, H.-Q. Khor, X. Feng, and G. Zhao, "Revealing the invisible with model and data shrinking for composite-database micro-expression recognition," *IEEE TIP*, vol. 29, 2020. 6

[39] L. Zhou, Q. Mao, X. Huang, F. Zhang, and Z. Zhang, "Feature refinement: An expression-specific feature learning and fusion method for micro-expression recognition," *Pattern Recognition*, vol. 122, 2022. 6

[40] X. B. Nguyen, C. N. Duong, X. Li, S. Gauch, H. S. Seo, and K. Luu, "Micron-bert: Bert-based facial micro-expression recognition," in *Proceedings of the IEEE/CVF CVPR*, June 2023. 6, 7

[41] Z. Zhai, J. Zhao, C. Long, W. Xu, S. He, and H. Zhao, "Feature representation learning with adaptive displacement generation and transformer fusion for micro-expression recognition," in *Proceedings of the IEEE/CVF CVPR*, June 2023. 6

[42] M. Verma, P. Lubal, S. K. Vipparthi, and M. Abdel-Mottaleb, "Rnas-mer: A refined neural architecture search with hybrid spatiotemporal operations for micro-expression recognition," in *Proceedings of the IEEE/CVF WACV*, January 2023. 6

[43] Z. Xie, J. Fan, and S. Cheng, "Multi-channel capsule network for micro-expression recognition with multiscale fusion," *Multimedia Tools and Applications*, Feb 2024. 6

[44] H. Zhang, L. Yin, H. Zhang, and X. Wu, "Facial micro-expression recognition using three-stream vision transformer network with sparse sampling and relabeling," *Signal, Image and Video Processing*, vol. 18, no. 4, pp. 3761–3771, Jun 2024. 6

[45] B. Xia, W. Wang, S. Wang, and E. Chen, *Learning from Macro-Expression: A Micro-Expression Recognition Framework*. Proceedings of the 28th ACM International Conference on Multimedia, 2020. 7

[46] Y. Li, X. Huang, and G. Zhao, "Joint local and global information learning with single apex frame detection for micro-expression recognition," *IEEE Transactions on Image Processing*, vol. 30, 2021. 7

[47] Y. Su, J. Zhang, J. Liu, and G. Zhai, "Key facial components guided micro-expression recognition based on first amp; second-order motion," in *IEEE International Conference on Multimedia and Expo*, 2021. 7

[48] X. Nie, M. A. Takalkar, M. Duan, H. Zhang, and M. Xu, "Geme: Dual-stream multi-task gender-based micro-expression recognition," *Neurocomputing*, vol. 427, 2021. 7

[49] L. Lei, J. Li, T. Chen, and S. Li, "A Novel Graph-TCN with a graph structured representation for micro-expression recognition," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020. 7

[50] W. Feng, M. Xu, Y. Chen, X. Wang, J. Guo, L. Dai, N. Wang, X. Zuo, and X. Li, "Nonlinear deep subspace network for micro-expression recognition," in *Proceedings of the 3rd Workshop on Facial Micro-Expression: Advanced Techniques for Multi-Modal Facial Expression Analysis*, 2023. 7